Why is that nucleotide there? Causes and consequences of kinetic and metabolic translational efficiency

Charneski, Catherine

Award date:
2014

Awarding institution:
University of Bath

Why is that nucleotide there? Causes and consequences of kinetic and metabolic translational efficiency

Catherine A. Charneski

A thesis submitted for the degree of Doctor of Philosophy
University of Bath
Department of Biology & Biochemistry
January 2014

Table of Contents

Acknowledgments

I thank Laurence for his scientific knowledge, opinion, and advice, all of which have always come with an unmatched ability to listen (rather than telling me to sod off as he might rightly have done). I could not have wanted a better supervisor, and I consider myself particularly fortunate in this regard. Much gratitude as well is owed for his graciously indulging my sometimes nuttiness and neuroticism, as well as his not-so-covert campaign to make me live more sustainably (see also: thanks for the loan of the entire Hurst multimedia collection). I will miss lunches!

The collective those who shall remain nameless will be fondly remembered for doorbells, faxes, the physics of everything (knowledge and porridge), as well as the ultimately affirmative task of querying the liking of various things. I thank my parents who have supported me and my friends who have always stood by me. Ostatni, ale nie mniej ważny, dzięki za darmowe przejazdy.

Contributions

Unless otherwise stated, all analyses were done by myself and interpreted together with my supervisor Laurence D. Hurst. In Chapter V, Frank Honti independently recapitulated some of the analyses in the main text that I performed as well, namely the separation of skew into differentially conserved sites, stop codon simulations, the correlation of amino acid usage with metabolic costs, and also contributed Supplemental Figures 9 and 10. All other investigations in Chapter V are my own.

Abbreviations

aa-tRNA	*aminoacyl-tRNA*
gespi	*gene strand preference index*
GFP	*green fluorescent protein*
GMP-PNP	*guanylyl imidodiphosphate*
HMM	*hidden Markov model*
PARS	*parallel analysis of RNA structure*
SNP	*single nucleotide polymorphism*
tAI	*tRNA adaptation index*
UTR	*untranslated region*

Summary

A variety of sequence features at the levels of DNA, RNA, and protein affect the process of translation and may thus be under selection to increase the efficiency of that process. To ask whether they are and in what ways they might act, I first examine what the sequence-based mechanisms of elongation rate determination are. I show that positive charges in newly-translated peptides are the primary sequence-encoded elements modulating elongation rates, probably via their interaction with the negatively charged ribosomal exit tunnel. Contrary to common expectations, I do not find that codon usage has a significant effect on the velocity of ribosomes under normal in vivo conditions, while mRNA structure has only a marginal effect. That codons do not significantly slow ribosomes compared to the magnitude of the charge effect is seemingly at odds with a large body of literature which purports to show that they in fact do. Reviewing the literature, however, I suggest that these apparently at-odds findings can be reconciled by considering the supply (available tRNA) and demand (transcriptomic codon usage plus translation initiation rates). Taking supply and demand into account reveals that if codons do slow ribosomes, they are likely to do so significantly only under highly non-equilibrium, experimental conditions. That codons may not greatly differ in their translation speeds one to the next under normal in vivo conditions calls into question theories of selection on codon usage bias to modulate translational efficiency in a local, along-transcript fashion, for example the suggestion that codon usage is selected for at the 5' end of transcripts as a kind of speed ramp to modulate ribosomal traffic just after translation initiation. That codons have similar translation speeds is still, however, consistent with a theory that codon usage is under general selection and even speed selection to increase the global translational efficiency of cells by limiting the number of bound ribosomes on mRNAs. Returning to the matter of positive charges, I ask whether they, instead of codons, might cause the suggested translational-ramping effect, as positive charges tend to be overloaded at the N-termini of proteins across various domains and taxa. I find however that their distribution at the starts of proteins is better explained by a biochemical, structural null rather than the gene regulatory hypothesis of the ramp: the use of positive charges at N-termini can be completely explained in terms of the needs of a subset of proteins to correctly orientate themselves in membranes. I end with an example of how selection for translational efficiency can act not just on the process of translation but on the finished protein product, showing that the need to manufacture metabolically cheap proteins contributes to the anomalous AT skews observed in the Firmicutes.

I. Introduction

That natural selection acts on phenotypes that confer differential fitness is well known. But behind the seeming simplicity of this statement lies a great deal of complexity in understanding how genomes and evolution work. What can constitute a phenotype? What can natural selection act on, and what are the causes of that variation?

The earliest examples of phenotypic changes that selection might act upon focused on observable organismal traits. These are the classic case studies of adaptation such as selection on beak size of Darwin's Galapagos finches (Lack 1947), or environmentally-responsive pigment changes in the peppered moth *Biston betularia* (Kettlewell 1955). As sequence data accumulated, signs of selection on gene coding sequences were found. There is evidence that a number of enzyme polymorphisms are advantageous, such that functionally significant changes in amino acid content at selected positions within proteins are selectively maintained within populations (e.g. Livingstone 1971; Clarke 1975; Hudson et al. 1994). A classic example is the well-known single amino acid polymorphism in β-hemoglobin (Ingram 1957), which, although lethal in double recessive form and conducive to sickle cell anaemia when heterozygous, undergoes balancing selection in some populations due to its protective effect against malaria (Currat et al. 2002). Thus, from the beginning, the earliest examples of selection applied to the most easily observable traits, or the physical products of genes themselves.

As both molecular data and an understanding of genomes grew, examples accumulated of how natural selection can act not just on the product but on the process of gene expression. Whether or not selection on the process occurs is non-trivial, as changes to processes brought about by changes in sequence can have knock-on consequences for the function of the gene product. One striking example is the proposition that the major phenotypic differences between chimp and human are not due to coding sequence alone. An initial investigation of a number of the two species' proteins showed they were surprisingly similar, and it was therefore suggested that much of phenotypic difference between the two species might be brought about at the regulatory level (King and Wilson 1975). This conjecture has at least to some extent been borne out by subsequent research. For example, many non-coding regions in humans which have undergone accelerated change since our divergence from chimpanzees are enhancers of developmental gene expression (Capra et al. 2013) and differences between DNA methylation in the two species often affect promoter activity and alternative splicing patterns (Fukuda et al. 2013). More generally, there are ample accounts of the evolution of *cis*-regulatory elements such as locally-acting promoters as well as long-range enhancers and insulators (Bartkuhn and Renkawitz 2008). Still further are examples of how selection can act, and has acted, on coding sequences themselves in

order to regulate the process of gene expression: *trans*-regulatory elements such as chaperones aid proper protein maturation (Beissinger and Buchner 1998); exonic splice enhancers affect the splicing and function of the eventual protein (Blencowe 2000); selection on mRNA helps dictate rates of transcript degradation (e.g. Su et al. 2007); and selection on coding sequences occurs to create histone binding sites (Warnecke et al. 2008). In this thesis I will present several analyses, all of which touch on what genomic sequence traces reveal about the process and/or products of gene expression.

THE VELOCITY OF TRANSLATION

I start by focusing on one case study of a biological process - that of translational velocity. Translation is a focal process in an actively dividing cell. In log-phase yeast, roughly 60% of transcription is directed towards the production of ribosomes, and 50% of RNA polymerase II activity and 90% of mRNA splicing act on ribosomal proteins (Warner 1999). During such intensive growth, about 13,000 (metabolically expensive) proteins are produced in a fast-growing yeast cell in one second alone (von der Haar 2008), with polypeptide elongation estimated at 10-20 amino acids per second in *Escherichia coli* (Young and Bremer 1976; Ruusala et al. 1984; Bremer and Dennis 1996). Such figures indicate a rough protein synthesis time of 17-33 seconds for a 1 kb gene. Although elongation rates are typically assumed to be slower in eukaryotes (Mathews et al. 2000), reported values span a wide range from 2 to 7 (Ross and Orlowski 1982), 3 to 10 (Boehlke and Friesen 1975), 9 (Waldron et al. 1977), or even 12 to 17 (Alberghina et al. 1975) amino acids polymerized per second. Additionally, many of these eukaryotic studies were carried out at lower temperatures (in particular Ross and Orlowski carried out their experiments at 22^{o}C), while the complementary prokaryotic studies of elongation rate were typically done at 37^{o}C (Milo 2013). Hence it remains unclear whether these decreased elongation rates in eukaryotes reflect true capacity differences or experimental alteration of enzyme catalytic efficiency via temperature (Milo 2013). Nonetheless, turnover rates of most proteins are slower than the rate of cell division (Larrabee et al. 1980). In line with this, translation is generally considered as a short-lived activity whose primary function is simply to decipher the genetic code in an mRNA into the corresponding amino acid sequence, thereby producing much longer-lived proteins.

Recently, however, the velocity of translation elongation has been implicated to have a more lasting effect on the outcome of translation than previously considered. Early studies investigating the length distributions of nascent peptide chains established that translation speed is not constant along the length of a given transcript (Protzel and Morris 1974; Chaney and Morris 1979; Lizardi et al. 1979; Randall et al. 1980; Varenne et al. 1982). Changes in the elongation speed of a ribosome along a single transcript have been shown to be capable of

influencing not just the general production but also the ultimate conformation, positioning, and functioning of a protein. For example, speeding elongation via a ribosomal mutation can hinder the ability of the firefly luciferase reporter to undergo proper co-translational folding, diminishing its specific activity (Siller et al. 2010). Regulation of translation elongation has also been shown to alter the final subcellular localization of proteins. Localization elements within the coding sequence of *ASH1* mRNA, for example, help ensure asymmetric sorting by stalling the completion of translation until the mRNA is correctly localized to the daughter cell, lest the translated protein is released from the ribosomal complex into the cytoplasm prematurely (Chartrand et al. 2002); this partitioning of this mRNA from budding yeast mother to daughter cells is essential to the production of the phenotype whereby mating type switching is repressed in the daughter cells specifically (Bobola et al. 1996). Similarly, another study demonstrated that translation termination of an ER tail-anchored protein is delayed in order to provide time for the correct chaperone to bind the protein and relay it to the endoplasmic reticulum for insertion (thus preventing its improper release into the cytosol) (Mariappan et al. 2010). Complete stalls in translation elongation can also have severe effects. Stalled ribosomes on transcripts function as error signals leading to degradation of those transcripts, a process termed 'no-go decay' (Doma and Parker 2006). Likewise, stalled ribosomes are implicated in protein degradation as they have been shown to trigger the decay of the unfinished nascent peptide chains associated with stalled ribosomal complexes (Brandman et al. 2012).

By what mechanism might translational velocity be altered? A major theme that I address in this thesis is the common consideration that selection on codons occurs to modulate the process, or speed of elongation. Before I introduce the literature behind this idea, I will first address what biased codon usage is, and how it is thought to come about.

What determines codon usage?

Most typically there are 20 amino acids coded for within any given genome. However the number of triplet codons that can exist given the four nucleotides that comprise DNA (4^3 or 64 codons) is far in excess of this, even accounting for the three standard stop codons. This is because the genetic code is redundant, i.e. each of the 20 typical amino acids can be coded for by more than one triplet codon in the standard code (as proposed by Crick (1958)). Such blocks of codons specifying the same amino acid are called synonymous codons. Codon usage within a given genome is non-random such that all codons within a synonymous block are not used with equal frequency (Fiers et al. 1971; Grantham et al. 1980; Ikemura 1981a; Bennetzen and Hall 1982; Gouy and Gautier 1982; Ikemura 1982). From organism to organism the so-called preferred or most common synonymous codon for a given amino acid may vary, such that the catalog of

preferred codons within a genome is species-specific (Grantham et al. 1980; Grantham et al. 1981; Sharp et al. 1988). The common wisdom is that codons which form a non-wobble (i.e. standard Watson-Crick pairing) at their third site with the anticodon are preferred (Grosjean and Fiers 1982). This trend however applies mainly to two-codon blocks, as four-codon blocks across different genomes display more erratic pairing patterns (Ran and Higgs 2010). Additionally, the vast number of identified epigenetic modifications to tRNAs that can restrict or expand codon recognition make overall wobble rules less than clear (Agris 2004).

In a given genome, why are some codons preferred at all? The first explanation for this historically was that of mutation and drift. If a mutation is neutral, it will not be seen by selection and its chance of fixation in the population is equal to its initial frequency in the population (Freese 1962; Sueoka 1962; Kimura 1968a, b). Early on synonymous mutations were indeed thought to be completely neutral as they do not change the amino acid sequence of the protein, and therefore no discernable effect on protein structure or function was to be expected (King and Jukes 1969). Even if a synonymous mutation had only a slight fitness consequence it may be carried within a population unseen by selection, particularly if that population has a small sample size (Ohta 1973). That synonymous mutations might be neutral or nearly neutral in terms of their fitness consequences was supported by observations that changes in synonymous third sites, which do not change the encoded protein and are therefore typically considered less functionally relevant than non-synonymous first and second codon sites (Sonneborn 1965), are more frequently observed than mutations in non-synonymous sites (Kimura 1977). Additional support was lent by reports that in mammals, substitutions at degenerate third sites do not occur any less frequently than in pseudogenes (Wolfe et al. 1989). Indeed, that selection is often thought to act on the gene product, or protein sequence, is still reflected in one of the major methods traditionally used to infer selection, the *K*a/*K*s ratio (Li and Gojobori 1983; Nei and Gojobori 1986). *K*a/*K*s is calculated as the ratio of the number of nonsynonymous substitutions per nonsynonymous site (*K*a) to the number of synonymous substitutions per synonymous site (*K*s) (Hurst 2002). In inferring selection via this ratio, synonymous substitutions are effectively taken to be neutral (Hurst 2002).

The fact that genomic GC content correlates with genomic codon usage across organisms was interpreted as further evidence that biased mutation pressure may have caused the divergence in codon usage profiles between species (Muto and Osawa 1987; Kanaya et al. 2001a; Knight et al. 2001). This is because variation in genomic GC content across species is typically assumed to reflect differences in mutation biases across taxa (Sueoka 1962). This assumption was buoyed by the observation that GC content varies widely among organisms (Barbu et al. 1956), yet within a given (prokaryotic) species there is comparatively little intragenomic heterogeneity in GC content

(Sueoka 1959). Correspondingly, in many archaeal and bacterial species, codon bias can be predicted by knowledge of the intergenic GC content alone (Chen et al. 2004). As GC content is typically assumed to be under weaker selective constraint and more indicative of mutational biases than coding regions, the ability to predict codon usage from nucleotide non-coding nucleotide content has been taken by some to be evidence that codon usage is primarily mutational in origin (Chen et al. 2004). However, as GC content in genomes has been recently found to be under selective constraint (Hershberg and Petrov 2010; Hildebrand et al. 2010), the relevance of inferences relying upon the assumption that GC content reflects mutation is arguably unclear.

Mutation coupled with drift is, nonetheless, not enough to explain the unequal codon usage observed in the bulk of sequenced genomes. There are indications that, in addition to mutational pressures acting on synonymous third sites, selection is acting within genomes to bias codon usage towards so-called "preferred" codons. First, mutation cannot explain why codon usage correlates with the tRNAs available to decode them in a variety of organisms from bacteria to humans (Post et al. 1979; Ikemura 1981a, 1982, 1985; Moriyama and Powell 1997; Percudani et al. 1997; Duret and Mouchiroud 1999; Kanaya et al. 1999; Kanaya et al. 2001b; dos Reis et al. 2004; Novoa et al. 2012). This tRNA:codon co-adaptation has also been detected in the particular expression profiles of differentiated metazoan cells, including both the silk gland of the silkworm *Bombyx mori* and rabbit reticulocytes (Chavancy et al. 1979) as well as within some human tissues (Dittmar et al. 2006; Waldman et al. 2010). Second, mutation cannot explain why more highly expressed genes often display much stronger codon usage bias (Post et al. 1979; Post and Nomura 1980; Gouy and Gautier 1982; Konigsberg and Godson 1983; Blake and Hinds 1984; Sharp and Li 1986; Dong et al. 1996). Nor can mutation bias explain why the codon usage in the most highly expressed genes matches the tRNA isoacceptor profile in a given genome, as was shown to be the case in a majority of sequenced microorganisms (Sharp et al. 2005). (I here single out two organisms of particular relevance to the analyses presented in this thesis: highly expressed genes in both *S. cerevisiae* and *E. coli* do display a bias toward codons which match the major tRNA species (Ikemura 1981b; Bennetzen and Hall 1982)). Third, fast growing bacteria display both greater codon usage bias in highly expressed genes and a stronger codon:tRNA co-adaptation (Rocha 2004). Finally, analyses of single base changes support the view that codon usage is selected. The stronger codon usage bias observed in highly expressed genes is often accompanied by a concomitant decrease in the substitution rate at silent (synonymous) sites, indicative of strong selection for codon usage bias (Ikemura 1985; Sharp and Li 1987; Sharp 1991) (although see also Eyre-Walker and Bulmer 1995). Additionally, in some species such as *Drosophila*, mutation bias is toward A+T, whereas preferred codons are GC rich (Powell and Moriyama 1997). That not all amino acids display the same nucleotide bias in synonymous third

sites for their preferred codons can also be looked upon as further evidence against the mutation argument (Powell and Moriyama 1997).

There are reports of genomes where selection does not appear to favor any synonymous codons (Lafay et al. 1999; Lafay et al. 2000; Sharp et al. 2005) – particularly in organisms with small population sizes or where the genomic composition is particularly GC or AT rich (Wright and Bibb 1992; Andersson and Sharp 1996; McInerney 1997). Generally, however, the mutation-selection-drift hypothesis of codon usage prevails. In this view, codon usage is selected for in highly expressed genes but a combination of mutation and drift keep rarer, less optimal codons present in lowly-expressed genes which are less subject to selection (Sharp and Li 1986; Bulmer 1991; Sharp et al. 1993; Akashi 1995; Knight et al. 2001; Rocha 2004).

Evidence for slow codons

Initial suggestions that codons might slow ribosomes were theoretical in nature. The first proposal that tRNA availability could modulate elongation rates postulated that the ribosome could slow if the availability of tRNA cognate to the A-site codon was low (Itano 1963). Further proposals along this theme followed. For example, it was hypothesized that the encoding of genes in the histidine operon by codons cognate to either major or minor tRNAs (of greater or lesser abundance, respectively) could dictate differences in the steady state levels of proteins whose genes lie within the same operon, a theory called "modulation control" (Ames and Hartman 1963; Stent 1964). A few years later, a theory of codon involvement in gene expression was developed wherein the role of tRNAs was to facilitate, rather than hinder, gene expression. Investigations of fibroin synthesis in the silk gland of *Bombyx mori* (silkworms) showed a correlation between the amino acid composition of fibroin protein and the corresponding aminoacylated tRNAs (aa-tRNAs) during the developmental phase when fibroin is secreted by the gland (Garel et al. 1970). This was proposed to be a functional adaptation to facilitate mass synthesis of a single protein by highly specialized cells, which Garel called the "tRNA adaptation theory" (1970).

These proposals were followed by reports that tRNA-mediated attenuation control systems have evolved in highly specific regulatory capacities, particularly in the case of a number of amino acid biosynthetic operons (Kolter and Yanofsky 1982). Such attenuator systems generally function via a 5' leader sequence which is enriched in codons coding for the amino acid which the enzymes coded for by the operon also produce. If charged tRNAs for these codons are present, ribosomes have little wait time and readily speed over the leader sequence. This allows the formation of a hairpin terminator structure which precludes transcription of the operon by RNA polymerase. If,

rather, cellular supplies of the focal amino acid are low, the concentration of aa-tRNAs correspondingly diminishes, causing the ribosome to lag at these codons. By occluding the mRNA locally, these slow ribosomes prevent formation of the terminator structure, and hence the genes coding for the much-needed amino acid anabolizing enzymes will be transcribed (Kolter and Yanofsky 1982). The system thus functions as a highly conserved auto-regulatory switch at specific loci.

The idea that all codons in a genome could affect translational efficiency, not just at specific regulatory loci or within specialized cells, was prompted when tRNA:codon co-adaptation was noted amongst a bulk of genomic coding sequences in *E. coli* (Post et al. 1979; Post and Nomura 1980; Ikemura 1981a). It was postulated that, considering the entire *E. coli* genome, codons corresponding to rare tRNAs would wait longer for the correct tRNA to enter the ribosomal active site, thus slowing translation over the aminoacylated-tRNA limited codons (Lizardi et al. 1979; Gouy and Gautier 1982; Grosjean and Fiers 1982). This conjecture was supported by observations that altering tRNA levels modulated the duration of ribosomal pausing in an in vitro translation system (Lizardi et al. 1979). Further evidence mounted that tRNA levels could modulate in vivo ribosomal speeds: rare codons, i.e. codons which are both genomically rare and correspond to rare tRNA species, were found to be more slowly translated in altered codon constructs expressed from multicopy plasmids (Pedersen 1984; Robinson et al. 1984; Bonekamp et al. 1985). Using pulse-chase experiments, Pedersen showed that the translation time of highly-expressed constructs containing more rare codons took significantly longer than time required to translate ribosomal proteins rich in common codons (Pedersen 1984). Similarly, introduction of multiple rare codons into a transgene in *E. coli* was shown to reduce the maximum possible translation rate of the transgene when expressed at high levels (Robinson et al. 1984). In separate experiments, comparison of a theoretical model of translation to the experimental distribution of peptide intermediate lengths for *E. coli* colicin A and a handful of other proteins, it was inferred that 1) diffusion of aminoacylated tRNAs plus cofactors (EfTU and GTP) to the ribosomal A-site is the rate-limiting step in translation, and 2) translation rate is inversely proportional to tRNA concentrations (Varenne et al. 1984). Additionally, upon replacement of common codons in the *E. coli pyrE* leader peptide (involved in a translational attenuation system) with rare ones, it was inferred from reduced expression of the gene that the rare codons were slowly translated (Bonekamp et al. 1985). However, it should be mentioned the design of Bonekamp's experiment, which introduced a frameshift and altered the transcriptional-translational coupling the attenuator system relies upon, was subsequently criticized (Roesser and Yanofsky 1991).

Thus codons are selected - but why? Are codons selected to have differential elongation rates? That codon bias is so strong in highly expressed genes suggests that it has something to do with

translational efficiency. But slowing is not the only reason why certain codons might be preferred. Alternative evidence has been put up to support the notion that synonymous codons are under selection for translational accuracy (see e.g. Akashi 1994; Stoletzki and Eyre-Walker 2007; Warnecke and Hurst 2010). Although not necessarily mutually exclusive with the notion that synonymous codons might also differ in their elongation rates, it offers an alternative reason why codons might be selected. For the purposes of this thesis, however, I will concentrate on the argument that codons modulate chain elongation rates.

What are the primary determinants of ribosomal slowing in vivo?

Before any discussion of whether codons are selected to slow can take place, we must first ask if codons slow. If so, we must ask how much, especially in relation to other possible slowing mechanisms. I have thus far focused on codon-mediated slowing due to the predominance of this idea in the literature, but there are other possible determinants of translational velocity. For example, transcript folding is another mRNA sequence feature that is known to impact the activity of the ribosome. That RNA secondary structure might impede the ability of the ribosome to elongate relative to single-stranded RNA was first proposed as a theoretical conjecture (Adams et al. 1969). The idea that mRNA folding might slow ribosomes was given further credence when Chaney and Morris investigated the distribution of peptide chains produced upon translation of the MS2 phage RNA which codes for the viral coat protein. Correlations between peptide lengths and locations of RNA structure led them to propose that the need to open up the folded RNA slowed elongation (Chaney and Morris 1979). A study of individual ribosomes travelling along a single mRNA molecule also indicated a role for nucleic acid secondary structure in impeding ribosomal progression (Wen et al. 2008). It was conversely suggested, however, that the magnitude of any slowing effect is much reduced once the ribosome has successfully initiated and is elongating along the length of the mRNA molecule (Kozak 1986), and thus the average effect on translation velocity that RNA structure has globally within a cell is still unclear. Still other sequence-based factors have been found to alter ribosomal speed. Sequences which bear a strong resemblance to the anti-Shine-Dalgarno sequence in *E. coli* bind the complementary sequence present in the 16S ribosomal RNA, thereby slowing ribosomes, and are generally underrepresented amongst this organism's mRNAs (Li et al. 2012). And finally, more recently positive charges in recombinant peptides have been linked to translational pausing mediated by an electrostatic interaction of the cation in the elongating chain with the negatively charged exit tunnel (Lu and Deutsch 2008). Such a mechanism represents a fundamentally different way of thinking about slowing, as the encoded determinant of elongation rate in this case resides in the protein itself, not at a nucleotide level.

Thus there are multiple possible sequence-based slowers of ribosomes with varying degrees of evidence in their favor. Do they all act to slow elongation under in vivo conditions? If so, what are the relative contributions of each? Experimental footprinting data (Ingolia et al. 2009) allows investigation of the issue in a way previously not possible. This dataset effectively profiles the locations ("footprints") of all the elongating ribosomes on endogenous yeast transcripts. By assuming the density of footprints is inversely proportional to the rate of translation, it is possible to interrogate which sequence features contribute most to slowing along transcripts. Difficulties in determining ribosomal speeds arise from the fact that many of the possible sequence determinants of ribosomal velocity are overlapping in nature in the first place. This is because a single nucleotide at one site can potentially alter the sequence features at a number of other levels—DNA, RNA, and/or protein. For example, a synonymous third-site may affect both codon bias and mRNA folding. In a similar vein, effects which at first glance appear to be a consequence of particular amino acid usage must be controlled against potential effects of the underlying codon usage among the synonymous group(s) in question. In Chapter II, I undertake a comparative analysis of ribosomal profiling data within and between yeast genes in order to examine the independent contributions of codon usage, mRNA structure, and positive charges on ribosomal slowing under normal, in vivo conditions. I find that both mRNA features only have a marginal, if any, effect on ribosome velocity. Instead, by far, the greatest effect on slowing comes not from the mRNA but from the protein itself, with slowing detectable just downstream of encoded positively charged residues. This is consistent with a mechanism for slowing whereby the newly-added positively charged amino acid interacts with the negatively-charged ribosomal exit tunnel as the ribosome continues elongating downstream and the nascent protein travels down the length of the tunnel.

RECONCILING THAT CODONS DO NOT SLOW AGAINST CONTRARY REPORTS

Thus, despite the common notion that codons differ markedly in their translation speeds, I find no evidence that they do so. Whether this is due to a lack of statistical power owing to the resolution of the footprint dataset that we employ (Ingolia et al. 2009) is an open question. However it seems certain that if there are systematic differences in the translation speeds of individual codons, they are not of a magnitude anywhere near that of the slowing induced by even a single positive charge, whose signal is clear. Yet there is a pervasive notion throughout the literature that codons greatly differ in their elongation speeds (e.g. Thanaraj and Argos 1996; Kimchi-Sarfaty et al. 2007; Higgs and Ran 2008). This assumption stems back to early experimental reports that codons slow, some of which I have already mentioned (e.g. Pedersen 1984; Robinson et al. 1984; Varenne et al. 1984). How then to square the finding I present in Chapter II with the fact that some experiments may find an effect of codons on speed?

In Chapter III I turn to this question of why codons do not differentially slow ribosomal elongation despite the evidence commonly rallied in favor of the hypothesis that they do. I review the literature that is often cited in support of the claim that codons slow and suggest that these findings, at seeming odds with the result in Chapter II, can be readily reconciled by two main points of thought. Firstly, some evidence does not in fact robustly show that codons slow and/or is improperly controlled for given other covariates of elongation speed which are now known—a prime example being the effect of positive charges. Yet, there is still outstanding literature which claims a large, significant effect of synonymous codons on translation speed, which cannot be dismissed on account of incorrectly controlled or poorly designed experiments. In fact there may be no leverage to claim that their findings are necessarily incorrect.

These remaining reports that codons slow can be explained by considering the second point, that the conception of the slowing problem, as put forward in this segment of the literature, is incomplete. That is, much of the experimental evidence that codons slow is treated as a function of codon usage alone. However, that tRNA availability can affect translation rates can be readily demonstrated by a simple thought experiment: in any organism, genes encoding a codon uniquely requiring a tRNA with a concentration of 0 will have a ribosomal velocity of 0 at those codon positions and thereafter along the transcripts. Conversely, an organism with a finite number of a given tRNA but an infinite number of codons absolutely requiring that tRNA will have an effective velocity of 0 at those codon positions. Thus instead of asking whether simply codons alone slow, it is likely more appropriate to consider the entire system as a supply and demand issue. In this formulation, speed is a consequence of both codon usage and the availability of their corresponding tRNAs (as also suggested by Liljenstrom et al. 1985; Curran and Yarus 1989; Qian et al. 2012; Pechmann and Frydman 2013). Reports that experimental alterations of tRNA concentrations modulate the degree of ribosomal pausing (Anderson 1969; Lizardi et al. 1979) corroborate this view. In this manner it is possible to reconcile the evidence, seemingly at odds with the finding presented in Chapter II, that under certain experimental conditions rare codons may slow. That is, it may be possible for codons to slow translation if tRNA supply and codon demand are experimentally thrown out of kilter compared to their typical, evolved in vivo balances wherein codon usage is proportional to the concentrations of tRNAs that decode them.

BIASED N-TERMINAL POSITIVE CHARGE USAGE

In the classical biochemical way of thinking, selection is considered to act on the protein itself, not the process of making it. That is, many amino acids are selected within proteins to perform certain concrete, even essential structural roles within the finished protein. The importance of the

physiochemical properties of amino acids in determining the structure and function of the protein product is underscored by the observation that nonsynonymous changes account for approximately half of all known disease polymorphisms in humans (Ng and Henikoff 2006). Examples of selectable features of amino acids include the following (note this list is not intended to be exhaustive but rather indicative of how different physiochemical properties of amino acid side chains may effect different functions). Glycine is the smallest amino acid as it has only a hydrogen atom for a side chain, and for this reason tends to make proteins flexible where it is inserted (Betts and Russell 2003). Consequently, this residue may play a role as a hinge. For example, glycine has been shown to act as a molecular switch gate in neuronal Na+ channels (Zhao et al. 2004), and within active sites, it is thought to aid the local intramolecular flexibility required for catalysis (Yan and Sun 1997). Conversely, proline is normally extremely inflexible within proteins as its side chain is connected to the amino acid backbone not once but twice (Betts and Russell 2003). This inflexibility causes proline to induce kinks in protein folds such as alpha helices, which may account for the extreme paucity of proline in globular proteins compared to other amino acids (Chou and Fasman 1974; O'Neil and DeGrado 1990). This reduction in the degrees of freedom for protein packing induced by proline can however confer greater thermostability (Hall and Reed 1957; Watanabe et al. 1991). Additionally, the ability for proline to undergo *cis-/trans*-isomerization renders it useful as an adaptive switch separating different enzyme conformations—for example, proline isomerization is thought to drive the transition between the open and closed states of a neuronal ion channel (Lummis et al. 2005). Cysteine is another amino acid that can render proteins more stable, with disulfide bridges between two cysteine capable of decreasing the entropy of packing structures (Betts and Russell 2003).

Classes of amino acids with side chains possessing similar physiochemical properties may also have general effects on protein structure. Hydrophobic residues are well known to play roles in membrane proteins, as such residues both form van der Waals interactions with membrane lipids and shield the polar groups in the peptide backbone from the lipid bilayer (Lodish et al. 2000; Rath and Deber 2012), lowering the free energy of the protein's configuration within the membrane. Such residues are also thought to be important in the complicated and poorly understood process of protein folding, as their tendency to escape water causes sequestration of hydrophobic residues to the interior of globular proteins (Kauzmann 1959; Rose et al. 1985). This hydrophobic force has been proposed to help drive the correct packing structures of globular proteins (Rose and Roy 1980; Dill 1985, 1990). Another physiochemical property of amino acid side chains, negative charge, can be selected to mediate the binding of one molecule to a positive charge in another, for example the binding of a receptor to its substrate (e.g. Czajkowski et al. 1993). And in halophilic organisms which maintain osmolarity with the external environment via

the internal accumulation of high concentrations of potassium ions, extensive negative charge use plays a particularly important role in maintaining the solubility of many proteins (Ebel et al. 1999).

The physiochemical property of amino acids I wish to focus on in particular in this thesis however is that of positive charge, which plays a great variety of structural roles within proteins. Such charges can mediate intra-molecular protein folding, for example alpha-helical formation (Sivaramakrishnan et al. 2008), or alter within-protein conformational rigidity/flexibility (Szeltner and Polgar 1996). Positive charges are also known to mediate interactions between the proteins they reside in and RNA, other protein- or lipid-binding partners. For example, the sliding clamp of T4 bacteriophage DNA polymerase is lined with positive charges so that DNA, with its negatively charged backbone, can be threaded through it (Moarefi et al. 2000). Positive charges lining ion channels can also allow for selective transport of metabolites (Doyle et al. 1998). Cations can play essential roles in the process of transmembrane signaling (Kim et al. 2012) and within active sites, they may play catalytic roles (Harris and Turner 2002). In contrast to hydrophobic residues, charged residues are often (but not always) exposed on the surface of proteins in order to help maintain protein solubility (Perutz et al. 1965; Shaw et al. 2001). Additionally, positive charges help orientate proteins within membranes: such proteins will position themselves within lipid bilayers so that the excess positive charge just adjacent to a hydrophobic transmembrane domain will tend to lie within the cytosol (von Heijne and Gavel 1988). Note, the understanding of the structural roles that amino acids play within proteins is of course more complicated than presented here. For instance, although lysine is normally positively charged at physiological pH, lysine residues which are buried in the protein interior can have drastically different pK_as, and thus have different propensities to be ionized (Isom et al. 2011). This finding underscores the importance of local environmental context in interpreting the physiochemical properties and function of different amino acid side chains within a mature, folded protein.

While amino acid residues can be under strong selection to perform particular functions in finished protein products (such as those examples presented above), it would likely be a mistake to think that every residue in a protein is under selection to be that specific residue. As mentioned previously, neutral or nearly neutral mutations can accumulate within genomes due to the stochasticity of genetic drift (Kimura 1968a; Ohta 1973). If such mutations occur in nonsynonymous sites, amino acids in the translated protein can be altered, without necessarily causing the loss of organismal viability (e.g. Jukes and Bhushan 1986). Such reasoning is held up to explain the surprising prevalence of enzyme polymorphism detected in early gel electrophoresis experiments (Ingram 1957; Harris 1966; Hubby and Lewontin 1966), all of which

cannot possibly be adaptive (Kimura and Ohta 1971; Ohta 1974). Further examples come from microorganisms which experience strong differential mutational pressures on the leading and lagging strands. These mutational differences, thought to arise from mechanistic differences in the way the leading and lagging strands are replicated, are capable of dramatically altering the nucleotide content on one replicatory strand compared to the other (Lobry 1996a; Mrázek and Karlin 1998). I will discuss this phenomenon, called nucleotide skew, further in the next Chapter. However for now I simply note that biased nucleotide content accumulating in nonsynonymous sites due to neutral, stochastic drift can then cause the divergence of the amino acid content of proteins encoded on each strand compared to the other (Lobry 1997; Mackiewicz et al. 1999).

Are major determinants of ribosomal velocity under systematic, large-scale selection to do so?

That sequence features at transcript starts might be able to regulate gene expression has been suggested before in various forms. For instance, rare codons have been hypothesized to be enriched in the leader sequence of membrane protein genes in order to slow translation, thereby ensuring the start of the secretion process before translation finishes (Burns and Beacham 1985) – an idea which was some time later independently re-suggested (Zalucki et al. 2009). Rare codons at transcript starts were also hypothesized to have another regulatory role, that of controlling the overall (steady state) rate of protein synthesis (Chen and Inouye 1990). Again similarly yet independently, it was later suggested that 'slow codons' are clustered at the 5' end of transcripts to act in some way as regulators of gene expression (Clarke and Clark 2010). These proposals remain unproven, with some evidence appearing against their favor. Adams (1969) initially suggested that selection on RNA structure might constrain the use of certain codons within genes, rather than vice versa. In line with this, abnormal synonymous codon usage at 5' transcript ends has been shown to be a product of selection on mRNA structure related to facilitating translation initiation, rather than on codon usage per se (Gu et al. 2010; Bentele et al. 2013; Goodman et al. 2013) as previously predicted (Eyre-Walker and Bulmer 1993). The finding that codons do not play a role in gene expression regulation via slowing of translation at 5' transcript starts is consistent with the finding I present in Chapter II that non-optimal codons do not significantly slow elongation in the data set under investigation.

Although the classical notion is that selection on non-synonymous content is due to selection on the product, or the amino acid sequence of the protein, a recent hypothesis suggests that amino acids may be selected at the N-terminus to modulate the process of translation (Tuller et al. 2011). This hypothesis is at least consistent with the finding I present in Chapter II, that positively charged residues are the primary determinants of the elongation rate. The use of positive charge is known to increase, on average, approaching N-termini in *E. coli* (Berezovsky et

al. 1999). That is, positive charge—while not present in every protein encoded in the organism—is on average enriched at protein starts, decreasing monotonically over the first thirty or so amino acids of proteins. Tuller et al. (2011) proposed that the increased average charge of amino acid residues at N-termini acts as a kind of ribosomal speed bump or 'ramp' at the beginnings of proteins in *E. coli* and *S. cerevisiae*. A hypothetical local slowing of ribosomes (compared to downstream speeds) just after initiation has been twice independently proposed as an adaptation to space ribosomes apart from one another, thus preventing traffic jams at the beginning of transcripts (Mitarai et al. 2008; Tuller et al. 2010), but only Tuller et al. (2011) outline a role for charges.

An N-terminal speed ramp could have far-reaching consequences for translational efficiency, if in its absence certain transcripts would be more prone to ribosomal bottlenecks. Although the mechanistic consequences of two ribosomes colliding are unknown, blockages in translation can lead to degradation of not only the aborted protein (Dimitrova et al. 2009) but of the transcript as well (Sunohara et al. 2004). Stalled translation can also have knock-on effects on translational efficiency in a more global sense, as too many ribosomes sequestered on stalled transcripts will prevent translation initiation on other much-needed transcripts (Andersson and Kurland 1990; Gingold and Pilpel 2011). As an extension of the adaptive ramp hypothesis, it has been suggested that as the (average) 'ramp' length roughly corresponds to the length of peptide occluded by the exit tunnel, a ramp of slow translation may exist at the beginning of proteins to facilitate their interactions with chaperones (Fredrick and Ibba 2010). Chaperones are known to bind subclasses of proteins to aid their correct folding, thereby aiding removal of potentially toxic, aggregation-prone misfolded intermediates (Beissinger and Buchner 1998; Hartl et al. 2011). It is, however, hard to see why translation would need to be slowed during the only time when any chaperones would be presumably unable to bind to the inaccessible nascent peptide, as the N-terminus of the newly synthesized chain would still be buried within the exit tunnel during the first thirty or so amino acids translated. These first thirty residues would then possibly emerge and only become accessible to chaperones during subsequent elongation proceeding at a more normal rate.

Could such charges be selected for in varied locations across proteins, in order to affect not just the final protein but the speed at which the ribosome travels down the mRNA in the vicinity of the charge residue, thereby regulating local translational processes? In Chapter IV I reflect on the idea of an adaptive ramp and examine the evidence in its favor. I note that the structural roles positive charges play within proteins were never incorporated into a null hypothesis of why average positive charge use increases at N-termini before the existence of such a translational ramp was claimed. That is, a reasonable null of positively charged amino acid use should incorporate the potential structural consequences of such charges: amino acids are the

constitutive building blocks of proteins, and reside within the body of the protein long after translation has finished. Re-investigating the issue, I find that the positive charge distribution at protein starts can indeed be entirely explained by biochemical forces, namely the need for proteins to orientate themselves in membranes via the 'positive inside' rule (Heijne 1986). This rule states that regions of positive charge adjacent to the membrane localize to the cytosol, thus helping dictate the topology of the protein within the lipid bilayer. The membrane orientation model makes a full account of the trend for average positive charge use to increase at N-termini; thus there is no need to invoke gene regulatory hypotheses to explain the distribution of N-terminal positive charges. This finding calls into doubt any involvement of these residues in modulating translational efficiency.

NUCLEOTIDE SKEW

I have thus far considered the arrangement of sequences on an amino acid level, within nascent proteins, and within gene coding sequences present in transcripts. I now wish to finish by taking a step back and investigating the arrangement of nucleotides in the DNA molecule as a whole, including both coding and non-coding sequences. In this Chapter, I examine a special case amongst bacterial genomes of how nucleotides are distributed within the two strands of a chromosome, and what that patterning reveals about underlying cellular processes.

The parity rules

As the significance of a pattern lies in its deviation from randomness, it is first necessary to ask what sort of null behavior can be expected regarding the distribution of nucleotides within DNA molecules. Watson and Crick (1953) discerned that the base pairing of pyrimidines with purines (A with T and G with C) in DNA allows two single yet complementary molecules of DNA to bind together into a double helix. Their discovery was based in part on the knowledge that both A and T nucleotides as well as G and C had been found to coexist in rough molar proportions, as determined by the enzymatic digestion of DNA (Chargaff 1950, 1951; Chargaff et al. 1952). Importantly, Chargaff's rule (that [A]~[T] and [G]~[C]; also known as parity rule 1) applies to the total nucleotide content of the DNA double helix considered as a whole. This equimolarity is a sheer consequence of the fact that generally only AT and GC base pairing are allowed between the two strands of duplex DNA. What, however, might we expect - if anything - about the relative proportions of different nucleotides within a single strand of the double helix? It was shown that there are certain conditions under which we can predict the relative frequencies of nucleotides within a single DNA strand (Lobry 1995; Sueoka 1995). Namely, if there is no bias in mutations or substitutions between the two complementary molecules, then the base

composition of the DNA can be predicted from six substitution types rather than the total twelve possible nucleotide substitutions (Sueoka 1995). Under such conditions, intra-strand substitution rates must be the same between the two strands, leading to the equilibrium situation where [A]~[T] and [G]~[C] within a single DNA molecule (parity rule 2) (Lobry 1995; Sueoka 1995).

There is some evidence to suggest that when considering a single strand to be one of the two complementary, physically distinct molecules of a chromosome, the second parity rule largely holds (although the reasons for this are not entirely clear) (Baisnée et al. 2002; Mitchell and Bridge 2006; Hart and Martínez 2011). That is, the relative proportions of complementary nucleotides within a single strand of DNA comprising one strand of a chromosome are often roughly equivalent. Still, it is arguably unclear whether such observations constitute a true proof of the second rule, as the genomes under consideration in the aforementioned studies (Baisnée et al. 2002; Hart and Martínez 2011) were not shown to be at compositional equilibrium. Nonetheless, and more relevant to the work presented in this thesis, the expectation that the second parity rule lays out is clearly violated with great frequency if the definition of the two strands is slightly modified. In the vast majority of bacterial genomes, an excess of one nucleotide over its complementary partner is observed between the two strands of replication; most commonly, this compositional asymmetry manifests as an excess of G over C (or positive GC skew, (G-C)/(G+C)) and to a lesser extent, T over A (negative AT skew, (A-T)/(A+T)) in the leading strand (Lobry 1996a, b; Blattner et al. 1997; Fraser et al. 1997; Kunst et al. 1997; Andersson et al. 1998; Grigoriev 1998; McLean et al. 1998; Mrázek and Karlin 1998; Tillier and Collins 2000; Lobry and Sueoka 2002).

Ways in which deviations from the second parity rule arise

Why are complementary nucleotides so often asymmetrically distributed within bacterial genomes? Nucleotide skews are generally thought to be mutational in nature due to the fact that they are strongest in sites thought to be subject to weaker selection, such as synonymous third sites and intergenic regions (Lobry 1996a; Blattner et al. 1997; Andersson et al. 1998; Tillier and Collins 2000; Lobry and Sueoka 2002). There is some contention, however, as to the mechanism producing these asymmetrical mutations. One prevalent explanation posits that genome-wide nucleotide skews result from differential mutational pressures incurred by the leading and lagging strand during replication (e.g. Wu and Maeda 1987; Lobry 1996a). This suggestion is supported by the observation that both GC and AT skews tend to switch their sign sharply at the origin and terminus of replication, suggesting a role for the replication fork in the generation of such skews (Lobry 1996a, b; Blattner et al. 1997; Fraser et al. 1997; Kunst et al. 1997; Andersson et al. 1998; Grigoriev 1998; McLean et al. 1998; Mrázek and Karlin 1998; Tillier and Collins 2000; Lobry and

Sueoka 2002). Differential mutational pressure is thought to result during replication on account of the asymmetry inherent to the DNA replication mechanism. The leading strand is synthesized with great processivity, but the lagging strand is synthesized discontinuously (Sakabe and Okazaki 1966). Indeed, a number of experiments have confirmed the increased mutability of lagging strand replication (Veaute and Fuchs 1993; Iwaki et al. 1996; Szczepanik et al. 2001) (although see also (Fijalkowska et al. 1998) who find that the leading strand replication is more mutagenic). The exact mutational profile which contributes to compositional asymmetry between the two strands remains unclear, although increased rates of cytosine deamination (C→T) in the lagging strand, which spends more time in a single-stranded state and is hence more vulnerable to this mutation type, is likely to play a role (Frank and Lobry 1999).

Other analyses nevertheless have implicated transcription-related mutation as the prime cause of skew (Francino and Ochman 1997; Nikolaou and Almirantis 2005). In this scenario the coding and non-coding strands, not the replicatory strands, are the division around which compositional asymmetry should evolve. Nucleotide skews could result during transcription either due to differential mutation and repair on the transcribed and non-transcribed strands. For example, increased rates of certain mutation types (e.g., cytosine deamination) may occur in the non-transcribed strand, which remains single-stranded and unprotected by RNA polymerase while the other strand is transcribed (Beletskii and Bhagwat 1996; Francino et al. 1996). Again, cytosine deamination may be at work, this time leading to an excess of C→T mutations in the coding strand, which would in turn result in an effective excess of G over C within the non-template strand (Francino and Ochman 1997). Additionally, only the transcribed strand will benefit from transcription-coupled repair mechanisms (Hanawalt 1989; Sweder and Hanawalt 1993; Beletskii and Bhagwat 1996), further biasing nucleotide asymmetries.

Replication- and transcription-based mutation ultimately, however, are not mutually exclusive explanations for nucleotide skew. In genomes with heavy loading of coding content onto the leading strand, it can be difficult to separate the two, as both replicational and transcriptional mutation and asymmetrical repair should be both biased to the leading strand. An ANOVA analysis which separated the contribution of these various factors found that replication- as opposed to transcription-based mutation explains more of the skew observed in selected bacterial genomes (Tillier and Collins 2000). It was conversely, however, also claimed that transcriptional effects are the greater contributor to skew (Touchon et al. 2004; Nikolaou and Almirantis 2005). Yet another study implicates both as playing significant contributions to skew (Necsulea and Lobry 2007). Thus the exact nature of the mechanism(s) whereby mutational processes cause genome-wide remains contentious, and it is possible that the answer may vary from organism to organism.

Do abnormal AT skews reflect unusual mutational processes?

There is a special case of compositional asymmetry I wish to address in this thesis - that of the Firmicutes. The gram-positive Firmicutes are anomalous among bacteria in that their leading strands display an excess of A over T (Morton and Morton 2007), rather than the classical T over A skew observed in other phyla (e.g. McLean et al. 1998). What do the unusual leading A>T skews in Firmicutes tell us about how genomic content and structure come about, and what implications might they have for how genomes function? Broadly, there are two possibilities underlying these atypical A>T skews. The first is that the Firmicutes display an unusual mutation bias compared to that seen in other bacteria. Is the process of replication or transcription in Firmicutes in fact different? The other alternative for the unexpected skews is that of selection. Firmicutes are unique in their extreme gene orientation bias (Rocha 2002; de Carvalho and Ferreira 2007). That is, in this group of bacteria there is a tendency for a remarkably high percentage (~78%) of their coding content to be encoded on the leading strand (Rocha 2002). Genome-wide nucleotide skews could result if there was sufficiently biased selection on coding content within open reading frames, purely on account of the asymmetric localization of genes between the two replicatory strands. In this case, possible selection on the gene product – for instance, selection acting on the amino acid content of proteins – that could produce the unusual chromosomal nucleotide skews. While such a selective explanation has been acknowledged to be possible in theory (Frank and Lobry 1999), one has never been shown.

Do mutational or selective processes shape skews in the Firmicutes? In order to answer this question, I start by dissecting the skew into different types of sites which are known to be under different types of selection. If skew is most predominant in third codon sites and intergenic regions, this might indicate a mutational origin. On the other hand, if only present in third sites but not intergenic regions, some form of translational selection on codon usage could be causing skew. Whether non-synonymous sites skew should indicate to what degree the pattern is caused by selection (on amino acid content) as opposed to mutation. By analyzing sites which are under different selective constraints, as above, I determine which sites skew the most. Using rare SNPs in one of the model species under question, *Staphylococcus aureus*, I am able to estimate the mutational equilibrium in putatively neutral sites to contrast with the observed skews. Via these intersecting approaches, I find that the Firmicutes display their anomalous skews not on account of a mutational bias, but because of selection on a type of translational efficiency different to that discussed thus far. That is, selection for amino acids which are metabolically cheap (require few ATP) to produce, coupled with the disparity in gene strandedness, cause the unusual A>T Firmicute skews. Thus, while I begin by showing that mechanistic effects of encoded components may slow but are not under a specific type of selection to do so, I end by showing

that selection on translational efficiency can contribute to the composition of the genomic landscape. The anomalous genome-wide nucleotide skews in the Firmicutes result from selection on the cost-efficiency of encoded gene products, not as a mutational by-product of essential genomic processes (replication and/or transcription).

References

Adams JM, Jeppesen PG, Sanger F, Barrell BG. 1969. Nucleotide sequence from the coat protein cistron of R17 bacteriophage RNA. Nature 223: 1009-1014.

Agris PF. 2004. Decoding the genome: a modified view. Nucleic Acids Res 32: 223-238.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. Genetics 139: 1067-1076.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics 136: 927-935.

Alberghina FA, Sturani E, Gohlke JR. 1975. Levels and rates of synthesis of ribosomal ribonucleic acid, transfer ribonucleic acid, and protein in Neurospora crassa in different steady states of growth. J Biol Chem 250: 4381-4388.

Ames BN, Hartman PE. 1963. The histidine operon. Cold Spring Harb. Symp. Quant. Biol. 28: 349.

Anderson WF. 1969. The effect of tRNA concentration on the rate of protein synthesis. Proc Natl Acad Sci U S A 62: 566-573.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. Microbiol Rev 54: 198-210.

Andersson SG, Sharp PM. 1996. Codon usage and base composition in Rickettsia prowazekii. J Mol Evol 42: 525-536.

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark UC, Podowski RM, Näslund AK, Eriksson AS, Winkler HH, Kurland CG. 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature 396: 133-140.

Baisnée PF, Hampson S, Baldi P. 2002. Why are complementary DNA strands symmetric? Bioinformatics 18: 1021-1033.

Barbu E, Lee KY, Wahl R. 1956. [Content of purine and pyrimidine base in deoxyribonucleic acid of bacteria]. Ann Inst Pasteur (Paris) 91: 212-224.

Bartkuhn M, Renkawitz R. 2008. Long range chromatin interactions involved in gene regulation. Biochim Biophys Acta 1783: 2161-2166.

Beissinger M, Buchner J. 1998. How chaperones fold proteins. Biol Chem 379: 245-259.

Beletskii A, Bhagwat AS. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. Proc Natl Acad Sci U S A 93: 13919-13924.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. J Biol Chem 257: 3026-3031.

Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol 9: 675.

Berezovsky IN, Kilosanidze GT, Tumanyan VG, Kisselev LL. 1999. Amino acid composition of protein termini are biased in different manners. Protein Eng 12: 23-30.

Betts MJ, Russell RB. 2003. Amino acid properties and consequences of substitutions. In: Barnes MR, Gray IC, editors. Bioinformatics for Geneticists: Wiley.

Blake RD, Hinds PW. 1984. Analysis of the codon bias in E. coli sequences. J Biomol Struct Dyn 2: 593-606.

Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. 1997. The complete genome sequence of Escherichia coli K-12. Science 277: 1453-1462.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25: 106-110.

Bobola N, Jansen RP, Shin TH, Nasmyth K. 1996. Asymmetric accumulation of Ash1p in postanaphase nuclei depends on a myosin and restricts yeast mating-type switching to mother cells. Cell 84: 699-709.

Boehlke KW, Friesen JD. 1975. Cellular content of ribonucleic acid and protein in Saccharomyces cerevisiae as a function of exponential growth rate: calculation of the apparent peptide chain elongation rate. J Bacteriol 121: 429-433.

Bonekamp F, Andersen HD, Christensen T, Jensen KF. 1985. Codon-defined ribosomal pausing in Escherichia coli detected by using the pyrE attenuator to probe the coupling between transcription and translation. Nucleic Acids Res 13: 4113-4123.

Brandman O, Stewart-Ornstein J, Wong D, Larson A, Williams CC, Li GW, Zhou S, King D, Shen PS, Weibezahn J, Dunn JG, Rouskin S, Inada T, Frost A, Weissman JS. 2012. A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. Cell 151: 1042-1054.

Bremer H, Dennis PP. 1996. Escherichia coli and Salmonella: Cellular and Molecular Biology. Washington DC: Am. Soc. Microbiol.

Bulmer M. 1991. The Selection-Mutation-Drift Theory of Synonymous Codon Usage. Genetics 129: 897-907.

Burns DM, Beacham IR. 1985. Rare codons in E. coli and S. typhimurium signal sequences. FEBS Lett 189: 318-324.

Capra JA, Erwin GD, McKinsey G, Rubenstein JL, Pollard KS. 2013. Many human accelerated regions are developmental enhancers. Philos Trans R Soc Lond B Biol Sci 368: 20130025.

Chaney WG, Morris AJ. 1979. Nonuniform size distribution of nascent peptides. The effect of messenger RNA structure upon the rate of translation. Arch Biochem Biophys 194: 283-291.

Chargaff E. 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia 6: 201-209.

Chargaff E. 1951. Structure and function of nucleic acids as cell constituents. Fed Proc 10: 654-659.

Chargaff E, Lipshitz R, Green C. 1952. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. J Biol Chem 195: 155-160.

Chartrand P, Meng XH, Huttelmaier S, Donato D, Singer RH. 2002. Asymmetric sorting of ash1p in yeast results from inhibition of translation by localization elements in the mRNA. Mol Cell 10: 1319-1330.

Chavancy G, Chevallier A, Fournier A, Garel JP. 1979. Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryote cell. Biochimie 61: 71-78.

Chen GF, Inouye M. 1990. Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the Escherichia coli genes. Nucleic Acids Res 18: 1465-1473.

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A 101: 3480-3485.

Chou PY, Fasman GD. 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13: 211-222.

Clarke B. 1975. The contribution of ecological genetics to evolutionary theory: detecting the direct effects of natural selection on particular polymorphic loci. Genetics 79 Suppl: 101-113.

Clarke TFt, Clark PL. 2010. Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. BMC Genomics 11: 118.

Crick FH. 1958. On protein synthesis. Symp Soc Exp Biol 12: 138-163.

Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J Mol Biol 209: 65-77.

Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, Clegg JB, Langaney A, Excoffier L. 2002. Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. Am J Hum Genet 70: 207-223.

Czajkowski C, Kaufmann C, Karlin A. 1993. Negatively charged amino acid residues in the nicotinic receptor delta subunit that contribute to the binding of acetylcholine. Proc Natl Acad Sci U S A 90: 6285-6289.

de Carvalho MO, Ferreira HB. 2007. Quantitative determination of gene strand bias in prokaryotic genomes. Genomics 90: 733-740.

Dill KA. 1990. Dominant forces in protein folding. Biochemistry 29: 7133-7155.

Dill KA. 1985. Theory for the folding and stability of globular proteins. Biochemistry 24: 1501-1509.

Dimitrova LN, Kuroha K, Tatematsu T, Inada T. 2009. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. J Biol Chem 284: 10343-10352.

Dittmar KA, Goodenbour JM, Pan T. 2006. Tissue-specific differences in human transfer RNA expression. PLoS Genet 2: e221.

Doma MK, Parker R. 2006. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. Nature 440: 561-564.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J Mol Biol 260: 649-663.

dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research 32: 5036-5044.

Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R. 1998. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science 280: 69-77.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proceedings Of the National Academy Of Sciences Of the United States Of America 96: 4482-4487.

Ebel C, Faou P, Franzetti B, Kernel B, Madern D, Pascu M, Pfister C, Richard S, Zaccai G. 1999. Molecular interactions in extreme halophiles - the solvation-stabilization hypothesis for halophilic proteins. In: Oren A, editor. Microbiology and biogeochemistry of hypersaline environments. Boca Raton, FL: CRC Press. p. 227-237.

Eyre-Walker A, Bulmer M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Research 21: 4599-4603.

Eyre-Walker A, Bulmer M. 1995. Synonymous substitution rates in enterobacteria. Genetics 140: 1407-1412.

Fiers W, Contreras R, De Wachter R, Haegeman G, Merregaert J, Jou WM, Vandenberghe A. 1971. Recent progress in the sequence determination of bacteriophage MS2 RNA. Biochimie 53: 495-506.

Fijalkowska IJ, Jonczyk P, Tkaczyk MM, Bialoskorska M, Schaaper RM. 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the Escherichia coli chromosome. Proc Natl Acad Sci U S A 95: 10020-10025.

Francino MP, Chao L, Riley MA, Ochman H. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272: 107-109.

Francino MP, Ochman H. 1997. Strand asymmetries in DNA evolution. Trends Genet 13: 240-245.

Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238: 65-77.

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, van Vugt R, Palmer N, Adams MD, Gocayne J, Weidman J, Utterback T, Watthey L, McDonald L, Artiach P, Bowman C, Garland S, Fuji C, Cotton MD, Horst K, Roberts K, Hatch B, Smith HO, Venter JC. 1997. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390: 580-586.

Fredrick K, Ibba M. 2010. How the sequence of a gene can tune its translation. Cell 141: 227-229.

Freese E. 1962. On the evolution of the base composition of DNA. J. Theor. Biol. 3: 82-101.

Fukuda K, Ichiyanagi K, Yamada Y, Go Y, Udono T, Wada S, Maeda T, Soejima H, Saitou N, Ito T, Sasaki H. 2013. Regional DNA methylation differences between humans and chimpanzees are associated with genetic changes, transcriptional divergence and disease genes. J Hum Genet 58: 446-454.

Garel JP, Mandel P, Chavancy G, Daillie J. 1970. Functional adaptation of tRNAs to fibroin biosynthesis in the silkgland of Bombyx mori L. FEBS Lett 7: 327-329.

Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. Mol Syst Biol 7: 481.

Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. Science 342: 475-479.

Gouy M, Gautier C. 1982. Codon usage in bacteria - correlation with gene expressivity. Nucleic Acids Research 10: 7055-7074.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9: r43-74.

Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8: r49-r62.

Grigoriev A. 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res 26: 2286-2290.

Grosjean H, Fiers W. 1982. Preferential codon usage in prokaryotic genes - the optimal codon anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18: 199-209.

Gu WJ, Zhou T, Wilke CO. 2010. A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. Plos Computational Biology 6.

Hall DA, Reed R. 1957. Hydroxyproline and thermal stability of collagen. Nature 180: 243.

Hanawalt PC. 1989. Preferential Repair of Damage in Actively Transcribed DNA- Sequences *in vivo*. Genome 31: 605-611.

Harris H. 1966. Enzyme polymorphisms in man. Proc R Soc Lond B Biol Sci 164: 298-310.

Harris TK, Turner GJ. 2002. Structural basis of perturbed pKa values of catalytic groups in enzyme active sites. IUBMB Life 53: 85-98.

Hart A, Martínez S. 2011. Statistical testing of Chargaff's second parity rule in bacterial genome sequences. Stochastic Models 27: 272-317.

Hartl FU, Bracher A, Hayer-Hartl M. 2011. Molecular chaperones in protein folding and proteostasis. Nature 475: 324-332.

Heijne G. 1986. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J 5: 3021-3027.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6.

Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol 25: 2279-2291.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6.

Hubby JL, Lewontin RC. 1966. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics 54: 577-594.

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. 1994. Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. Genetics 136: 1329-1340.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18: 486-487.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2: 13-34.

Ikemura T. 1981a. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146: 1-21.

Ikemura T. 1981b. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151: 389-409.

Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158: 573-597.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218-223.

Ingram VM. 1957. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature 180: 326-328.

Isom DG, Castaneda CA, Cannon BR, Garcia-Moreno B. 2011. Large shifts in pKa values of lysine residues buried inside a protein. Proc Natl Acad Sci U S A 108: 5260-5265.

Itano H editor. Proc. Symp. Abnormal Haemoglobins. 1963 Ibadan, Nigeria.

Iwaki T, Kawamura A, Ishino Y, Kohno K, Kano Y, Goshima N, Yara M, Furusawa M, Doi H, Imamoto F. 1996. Preferential replication-dependent mutagenesis in the lagging DNA strand in Escherichia coli. Mol Gen Genet 251: 657-664.

Jukes TH, Bhushan V. 1986. Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J Mol Evol 24: 39-44.

Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nishi T, Mori H, Ikemura T. 2001a. Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome. Gene 276: 89-99.

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001b. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol 53: 290-298.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238: 143-155.

Kauzmann W. 1959. Some factors in the interpretation of protein denaturation. Adv Protein Chem 14: 1-63.

Kettlewell HBD. 1955. Selection experiments on industrial melanism in the Lepidoptera. Heredity 9: 323-342.

Kim C, Schmidt T, Cho EG, Ye F, Ulmer TS, Ginsberg MH. 2012. Basic amino-acid side chains regulate transmembrane integrin signalling. Nature 481: 209-213.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315: 525-528.

Kimura M. 1968a. Evolutionary rate at the molecular level. Nature 217: 624-626.

Kimura M. 1968b. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res 11: 247-269.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275-276.

Kimura M, Ohta T. 1971. Protein polymorphism as a phase of molecular evolution. Nature 229: 467-469.

King JL, Jukes TH. 1969. Non-Darwinian evolution. Science 164: 788-798.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188: 107-116.

Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol 2: RESEARCH0010.

Kolter R, Yanofsky C. 1982. Attenuation in amino acid biosynthetic operons. Annu Rev Genet 16: 113-134.

Konigsberg W, Godson GN. 1983. Evidence for use of rare codons in the dnaG gene and other regulatory genes of Escherichia coli. Proc Natl Acad Sci U S A 80: 687-691.

Kozak M. 1986. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. Proc Natl Acad Sci U S A 83: 2850-2854.

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Codani JJ, Connerton IF, Danchin A, et al. 1997. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390: 249-256.

Lack D. 1947. Darwin's finches. Cambridge, UK: Cambridge University Press.

Lafay B, Atherton JC, Sharp PM. 2000. Absence of translationally selected synonymous codon usage bias in Helicobacter pylori. Microbiology 146 (Pt 4): 851-860.

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27: 1642-1649.

Larrabee KL, Phillips JO, Williams GJ, Larrabee AR. 1980. The relative rates of protein synthesis and degradation in a growing culture of Escherichia coli. J Biol Chem 255: 4125-4130.

Li W-H, Gojobori T. 1983. Rapid evolution of goat and sheep globin genes following gene duplication. Molecular Biology and Evolution 1: 94-108.

Liljenstrom H, von Heijne G, Blomberg C, Johansson J. 1985. The tRNA cycle and its relation to the rate of protein synthesis. Eur Biophys J 12: 115-119.

Livingstone FB. 1971. Malaria and human polymorphisms. Annu Rev Genet 5: 33-64.

Lizardi PM, Mahdavi V, Shields D, Candelas G. 1979. Discontinuous translation of silk fibroin in a reticulocyte cell-free system and in intact silk gland cells. Proc Natl Acad Sci U S A 76: 6211-6215.

Lobry JR. 1996a. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13: 660-665.

Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205: 309-316.

Lobry JR. 1996b. Origin of replication of Mycoplasma genitalium. Science 272: 745-746.

Lobry JR. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol 40: 326-330.

Lobry JR, Sueoka N. 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol 3: RESEARCH0058.

Lodish H, Berk A, Zipursky SL, al e. 2000. Section 3.4, Membrane Proteins. In. Molecular Cell Biology. New York: W. H. Freeman.

Lu J, Deutsch C. 2008. Electrostatics in the ribosomal tunnel modulate chain elongation rates. J Mol Biol 384: 73-86.

Lummis SC, Beene DL, Lee LW, Lester HA, Broadhurst RW, Dougherty DA. 2005. Cis-trans isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. Nature 438: 248-252.

Mackiewicz P, Gierlik A, Kowalczuk M, Dudek MR, Cebrat S. 1999. How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res 9: 409-416.

Mariappan M, Li X, Stefanovic S, Sharma A, Mateja A, Keenan RJ, Hegde RS. 2010. A ribosome-associating factor chaperones tail-anchored membrane proteins. Nature 466: 1120-1124.

Mathews MB, Sonenberg N, Hershey JWB. 2000. Origins and principles of translational control. In: Sonenberg N, Hershey JWB, Mathews MB, editors. Translational Control of Gene Expression. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. p. 1-31.

McInerney JO. 1997. Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb. Comp. Genomics 2: 1-10.

McLean MJ, Wolfe KH, Devine KM. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol 47: 691-696.

Milo, R. Cell biology by the numbers: What is faster, transcription or translation? [Internet]. 2013. Available from: http://www.weizmann.ac.il/plants/Milo/images/FasterTranscriptionTranslation100118Clean.pdf

Mitarai N, Sneppen K, Pedersen S. 2008. Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. J Mol Biol 382: 236-245.

Mitchell D, Bridge R. 2006. A test of Chargaff's second rule. Biochem Biophys Res Commun 340: 90-94.

Moarefi I, Jeruzalmi D, Turner J, O'Donnell M, Kuriyan J. 2000. Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. J Mol Biol 296: 1215-1223.

Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in *Drosophila*. Journal of Molecular Evolution 45: 514-523.

Morton RA, Morton BR. 2007. Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes. BMC Genomics 8: 369.

Mrázek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci U S A 95: 3720-3725.

Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proceedings Of the National Academy Of Sciences Of the United States Of America 84: 166-169.

Necsulea A, Lobry JR. 2007. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol Biol Evol 24: 2169-2179.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. Molecular Biology and Evolution 3: 418-426.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7: 61-80.

Nikolaou C, Almirantis Y. 2005. A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. Nucleic Acids Res 33: 6816-6822.

Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. 2012. A role for tRNA modifications in genome structure and codon usage. Cell 149: 202-213.

O'Neil KT, DeGrado WF. 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 250: 646-651.

Ohta T. 1974. Mutational pressure as the main cause of molecular evolution and polymorphism. Nature 252: 351-354.

Ohta T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. Nature 246: 96-98.

Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 20: 237-243.

Pedersen S. 1984. Escherichia coli ribosomes translate in vivo with variable rate. EMBO J 3: 2895-2898.

Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol 268: 322-330.

Perutz MF, Kendrew JC, Watson HC. 1965. Structure and function of haemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. J Mol Biol 13: 669-678.

Post LE, Nomura M. 1980. DNA sequences from the str operon of Escherichia coli. J Biol Chem 255: 4660-4666.

Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP. 1979. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl Acad Sci U S A 76: 1697-1701.

Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in *Drosophila*. Proceedings Of the National Academy Of Sciences Of the United States Of America 94: 7784-7790.

Protzel A, Morris AJ. 1974. Gel chromatographic analysis of nascent globin chains. Evidence of nonuniform size distribution. J Biol Chem 249: 4594-4600.

Qian W, Yang J-R, Pearson NM, Maclean C, Zhang J. 2012. Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. PLoS Genet 8: e1002603.

Ran W, Higgs PG. 2010. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol Biol Evol 27: 2129-2140.

Randall LL, Josefsson LG, Hardy SJ. 1980. Novel intermediates in the synthesis of maltose-binding protein in Escherichia coli. Eur J Biochem 107: 375-379.

Rath A, Deber CM. 2012. Protein structure in membrane domains. Annu Rev Biophys 41: 135-155.

Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican A, Eaton M, Humphreys G. 1984. Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res 12: 6663-6671.

Rocha E. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 10: 393-395.

Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14: 2279-2286.

Roesser JR, Yanofsky C. 1991. The effects of leader peptide sequence and length on attenuation control of the trp operon of E.coli. Nucleic Acids Res 19: 795-800.

Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. 1985. Hydrophobicity of amino acid residues in globular proteins. Science 229: 834-838.

Rose GD, Roy S. 1980. Hydrophobic basis of packing in globular proteins. Proc Natl Acad Sci U S A 77: 4643-4647.

Ross JF, Orlowski M. 1982. Growth-rate-dependent adjustment of ribosome function in chemostat-grown cells of the fungus Mucor racemosus. J Bacteriol 149: 650-653.

Ruusala T, Andersson D, Ehrenberg M, Kurland CG. 1984. Hyper-accurate ribosomes inhibit growth. EMBO J 3: 2575-2580.

Sakabe K, Okazaki R. 1966. A unique property of the replicating region of chromosomal DNA. Biochim Biophys Acta 129: 651-654.

Sharp PM. 1991. Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. J Mol Evol 33: 23-33.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucl. Acids Res. 33: 1141-1153.

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. 1988. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res 16: 8207-8211.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution 24: 28-38.

Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Molecular Biology and Evolution 4: 222-230.

Sharp PM, Stenico M, Peden JF, Lloyd AT. 1993. Codon usage - mutational bias, translational selection, or both. Biochemical Society Transactions 21: 835-841.

Shaw KL, Grimsley GR, Yakovlev GI, Makarov AA, Pace CN. 2001. The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. Protein Sci 10: 1206-1215.

Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM. 2010. Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. J Mol Biol 396: 1310-1318.

Sivaramakrishnan S, Spink BJ, Sim AY, Doniach S, Spudich JA. 2008. Dynamic charge interactions create surprising rigidity in the ER/K alpha-helical protein motif. Proc Natl Acad Sci U S A 105: 13356-13361.

Sonneborn TM. 1965. Degeneracy of the genetic code, extent, nature, and genetic implications. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. p. 377-397.

Stent GS. 1964. The Operon: On its third anniversary. Modulation of transfer RNA species can provide a workable model of an operator-less operon. Science 144: 816-820.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: Selection for translational accuracy. Molecular Biology and Evolution 24: 374-381.

Su YQ, Sugiura K, Woo Y, Wigglesworth K, Kamdar S, Affourtit J, Eppig JJ. 2007. Selective degradation of transcripts during meiotic maturation of mouse oocytes. Dev Biol 302: 104-117.

Sueoka N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40: 318-325.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A 48: 582-592.

Sueoka N. 1959. A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc Natl Acad Sci U S A 45: 1480-1490.

Sunohara T, Jojima K, Tagami H, Inada T, Aiba H. 2004. Ribosome stalling during translation elongation induces cleavage of mRNA being translated in Escherichia coli. J Biol Chem 279: 15368-15375.

Sweder KS, Hanawalt PC. 1993. Transcription-coupled DNA-repair. Science 262: 439-439.

Szczepanik D, Mackiewicz P, Kowalczuk M, Gierlik A, Nowicka A, Dudek MR, Cebrat S. 2001. Evolution rates of genes on leading and lagging DNA strands. J Mol Evol 52: 426-433.

Szeltner Z, Polgar L. 1996. Conformational stability and catalytic activity of HIV-1 protease are both enhanced at high salt concentration. J Biol Chem 271: 5458-5463.

Thanaraj TA, Argos P. 1996. Ribosome-mediated translational pause and protein domain organization. Protein Sci 5: 1594-1612.

Tillier ER, Collins RA. 2000. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol 50: 249-257.

Touchon M, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2004. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic Acids Res 32: 4969-4978.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141: 344-354.

Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12: R110.

Varenne S, Buc J, Lloubes R, Lazdunski C. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol 180: 549-576.

Varenne S, Knibiehler M, Cavard D, Morlon J, Lazdunski C. 1982. Variable rate of polypeptide chain elongation for colicins A, E2 and E3. J Mol Biol 159: 57-70.

Veaute X, Fuchs RP. 1993. Greater susceptibility to mutations in lagging strand of DNA replication in Escherichia coli than in leading strand. Science 261: 598-600.

von der Haar T. 2008. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst Biol 2: 87.

von Heijne G, Gavel Y. 1988. Topogenic signals in integral membrane proteins. Eur J Biochem 174: 671-678.

Waldman YY, Tuller T, Shlomi T, Sharan R, Ruppin E. 2010. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. Nucleic Acids Research 38: 2964-2974.

Waldron C, Jund R, Lacroute F. 1977. Evidence for a high proportion of inactive ribosomes in slow-growing yeast cells. Biochem J 168: 409-415.

Warnecke T, Batada NN, Hurst LD. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. PLoS Genet 4: e1000250.

Warnecke T, Hurst LD. 2010. GroEL dependency affects codon usage--support for a critical role of misfolding in gene evolution. Mol Syst Biol 6: 340.

Warner JR. 1999. The economics of ribosome biosynthesis in yeast. Trends Biochem Sci 24: 437-440.

Watanabe K, Chishiro K, Kitamura K, Suzuki Y. 1991. Proline residues responsible for thermostability occur with high frequency in the loop regions of an extremely thermostable oligo-1,6-glucosidase from Bacillus thermoglucosidasius KP1006. J Biol Chem 266: 24287-24294.

Watson JD, Crick FH. 1953. Genetical implications of the structure of deoxyribonucleic acid. Nature 171: 964-967.

Wen JD, Lancaster L, Hodges C, Zeri AC, Yoshimura SH, Noller HF, Bustamante C, Tinoco I. 2008. Following translation by single ribosomes one codon at a time. Nature 452: 598-603.

Wolfe KH, Sharp PM, Li W-H. 1989. Mutation rates differ among regions of the mammalian genome. Nature 337: 283-285.

Wright F, Bibb MJ. 1992. Codon usage in the G+C-rich Streptomyces genome. Gene 113: 55-65.

Wu CI, Maeda N. 1987. Inequality in mutation rates of the two strands of DNA. Nature 327: 169-170.

Yan BX, Sun YQ. 1997. Glycine residues provide flexibility for enzyme active sites. J Biol Chem 272: 3190-3194.

Young R, Bremer H. 1976. Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate. Biochem J 160: 185-194.

Zalucki YM, Beacham IR, Jennings MP. 2009. Biased codon usage in signal peptides: a role in protein export. Trends Microbiol 17: 146-150.

Zhao Y, Yarov-Yarovoy V, Scheuer T, Catterall WA. 2004. A gating hinge in Na+ channels; a molecular switch for electrical signaling. Neuron 41: 859-865.

II. *Positively charged residues are the major determinants of ribosomal velocity*

Catherine A. Charneski & Laurence D. Hurst

PLoS Biol (2013) 11(3): e1001508

OPEN ACCESS Freely available online

PLOS | BIOLOGY

Positively Charged Residues Are the Major Determinants of Ribosomal Velocity

Catherine A. Charneski, Laurence D. Hurst*

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

Abstract

Both for understanding mechanisms of disease and for the design of transgenes, it is important to understand the determinants of ribosome velocity, as changes in the rate of translation are important for protein folding, error attenuation, and localization. While there is great variation in ribosomal occupancy along even a single transcript, what determines a ribosome's occupancy is unclear. We examine this issue using data from a ribosomal footprinting assay in yeast. While codon usage is classically considered a major determinant, we find no evidence for this. By contrast, we find that positively charged amino acids greatly retard ribosomes downstream from where they are encoded, consistent with the suggestion that positively charged residues interact with the negatively charged ribosomal exit tunnel. Such slowing is independent of and greater than the average effect owing to mRNA folding. The effect of charged amino acids is additive, with ribosomal occupancy well-predicted by a linear fit to the density of positively charged residues. We thus expect that a translated poly-A tail, encoding for positively charged lysines regardless of the reading frame, would act as a sandtrap for the ribosome, consistent with experimental data.

Citation: Charneski CA, Hurst LD (2013) Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. PLoS Biol 11(3): e1001508. doi:10.1371/journal.pbio.1001508

Academic Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received June 19, 2012; **Accepted** February 1, 2013; **Published** March 12, 2013

Funding: LDH is a Wolfson Royal Society Research Merit Award Holder. CAC is funded by the University of Bath. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: PARS, parallel analysis of RNA structure; tAI, tRNA adaptation index.

* E-mail: L.D.Hurst@bath.ac.uk

Introduction

While it is known that there is great variation in ribosomal velocity along even a single transcript [1], what determines how fast a transcript (or part thereof) is processed is unresolved. Resolving this issue is important for understanding causes of disease and for the generation of transgenes, as changes in the local translation rate along mRNAs have been implicated in the regulation of protein folding [2], error attenuation processes such as no-go decay in yeast [3], transcription attenuation in bacterial systems [4], and correct protein localization [5,6].

For some time it has been hypothesized [7–10], and commonly assumed (e.g., [11,12]), that codons matching rare tRNAs slow ribosomes along transcripts due to differential tRNA availability. The supposition is that codons corresponding to less abundant tRNAs are translated at slower rates as the ribosome must pause while the appropriate tRNA becomes available. This, for example, is held up to explain the usage of codons specified by the most abundant tRNAs in the most highly expressed genes [13,14]. Although the notion that rare codons must stall ribosomes is commonplace, recent work has started to undermine the supposition that differential usage of synonymous codons will significantly alter the rate of ribosomal translocation within a transcript under normal conditions [15–17]. Indeed, much of the evidence cited as support for an effect on translational speed is questionable (see Note S1) and many of the patterns attributed to selection for translational speed are better explained in terms of selection on codon usage for translational accuracy [18–21].

Codon usage, however, is not the only potential factor affecting elongation speed. Double-stranded mRNA hairpin or pseudoknot structures are thought to impede progress of the ribosome [22,23]. The generality of this during elongation, however, is unclear, as other studies [24] suggest that the ribosome can more readily melt moderately stable secondary structures once initiation has taken place.

While the above factors consider ribosomal velocity to be modulated by properties of the mRNA, much less attention has been paid to the possibility that the resultant protein might impact translation rates. However, recent experimental work on recombinant peptides has shown that positive charges on the newly synthesized peptide might slow ribosomes [25,26]. This is conjectured to be owing to an electrostatic interaction between the cation in the emerging polypeptide and the negatively charged exit tunnel of the ribosome [25,26]. Following on from this, it has been suggested that positive charges, codon usage bias, and transcript folding play a role in ribosomal stalling at 5′ transcript ends [27,28].

Here we ask not whether certain features can sometimes modulate translation speed along a transcript (e.g., when grossly overrepresented in transgenes; see Note S1), but if they do as evolved in endogenous genes when expressed at "normal" levels, and to what extent. Ribosomally protected mRNA footprints from an experimental *Saccharomyces cerevisiae* dataset [29] enable us to profile the location of ribosomes across the *S. cerevisiae* transcriptome. Under the assumption that ribosomal densities inversely reflect ribosomal velocity [30,31], we independently examine the

Author Summary

Ribosomes do not synthesize protein at a constant rate along transcripts, and changes in translation speed can have knock-on consequences for the expression of that protein, even altering its folding or subcellular localization. It has long been thought that RNA-level features modulate translation rates, whether by delays incurred through the presence of codons that require relatively rare tRNAs, or by regions of mRNA folding that physically impede ribosomal progression. We find on the contrary that it is not RNA-level features but positive charges in the already translated protein that most retard ribosomes, possibly by interacting with the negatively charged ribosomal exit tunnel. We show that positive charge explains the sites where ribosomes stall most commonly within transcripts. We also show why, if protein charge were not considered, one could be misled into suspecting a role for non-optimal codons. Finally, we observe that the poly-A tail provides a massively positively charged terminus no matter in which frame it is translated. A missed stop codon or frameshifting would then lead to a stalled ribosome, which is consistent with experimental data.

effects of codon usage, mRNA folding, and positive charge on ribosomal speed throughout endogenous yeast genes. We show that positive charges in the nascent peptide slow the ribosome along transcripts in an additive manner in vivo, and that this slowing effect cannot be accounted for by mRNA structure, and even far surpasses that (if any) induced by codon usage bias. Within transcripts, those regions with the highest ribosomal occupancy are those most likely to be just downstream of positively charged residues. The cation sandtrap effect has potential relevance for the evolution of the poly-A tail, specifying as it does a series of positively charged amino acids if translated.

Results

While some recent work on nucleotide-resolution ribosomal footprint data [29] has claimed that codon usage plays a role in slowing ribosomes [27,28], another study that examined the same footprint data, filtered for noise, contradicts this claim [16]. Here we reanalyze the same dataset using both stringent mapping to reduce false-positive footprints (see Methods, "Ribosomal Density Data" for further comments on this and previous studies) as well as a novel normalization method to detect any accrual of ribosomal density, on average across transcripts, after putative ribosome-slowing features.

Neither Clusters of Nor Consecutive Rare Codons Tend to Slow Ribosomes

Ribosomal footprint data [29] allow us to examine changes in the rate of translation given the assumption that the slower a ribosome travels along a given portion of a transcript, the more likely it is to be found there at any point in time [30,31]. In the case of codon usage, we expect to see any possible ribosomal stalling centered over the rare codon(s) while the ribosome awaits a tRNA to enter its A-site. Hence to examine the effect of a sequence feature such as rare codons on the speed of translation, we calculate the relative change in stringently mapped ribosomal densities that occurs within a single transcript as ribosomes begin to translate regions of transcript enriched for rare codons (see Methods and Figure 1). To this end, within each transcript we compared the ribosomal occupancy at codon positions (r_{pos}) in the vicinity of clusters of rare codons (r_{pos}) to the average ribosomal occupancy of the 30 codons immediately preceding the first rare codon in the cluster (r_{prec30}). We then averaged the relative increase or decrease in ribosomal occupancy across transcript sections aligned by rare codon clusters. A mean r_{pos}/r_{prec30} after the clusters >1 indicates a denser sampling of ribosomal footprints on average and hence slowing at that codon position, while a mean $r_{pos}/r_{prec30}<1$ denotes sparser ribosomal coverage, consistent with acceleration.

In our main analysis we make use of the tAI (range 0–1) as a measure of codon optimality as this metric uniquely reflects the tRNA pool. The tAI of a sequence is defined as the geometric mean of the relative adaptiveness of its constituent codons to the tRNA pool available in that organism [32]. A higher tAI indicates the codon has a high abundance of decoding isoacceptor tRNAs and, according to the codon usage hypothesis of translational speed, should be translated faster on account of its ready coupling with an aminoacylated tRNA. A lower tAI conversely indicates a codon that is matched by a low number of tRNAs and is therefore putatively slowly translated and nonoptimal. Here we define "rare" codons to be those in the lowest quartile of tAI values (Methods, "The Average Effect of Codon Usage on Ribosomal Densities") (see also Figures S1, S2, S3 and Table S1 for analysis of rare codons defined according to genomic frequency).

Our results show inconsistent trends in ribosomal occupancy after rare codon clusters when all clusters of a given size are aligned and the average increase in ribosomal density after the cluster (here uncontrolled for covariates) is plotted (Figure 2A). This inconsistency is still apparent when we consider rare codons to be not those with a low tAI but those that are genomically infrequent (Figure S1). If there is any slowing due to rare codons, we should expect an increase in the amount of slowing along the mRNA as the number of rare codons increases. However, no such trend is evident (Figure 3A). This lack of influence of rare codon usage on ribosomal speed is not owing to a covariance between rare codon clusters and expression levels (Table S2). Shifting the location of the "preceding 30 codons" we use to normalize footprint values slightly upstream, to accommodate the 5′ portion of the ribosome potentially slowed over a rare codon, still detects no slowing due to codon usage (Figure S4).

As it has been postulated that tandem nonoptimal codons may more strongly inhibit progression of the ribosome than scattered rare codons [33,34], we also investigated whether consecutive rare codons (adjacent codons, each from the lowest quartile of tAI values) may be affecting ribosomal velocity. Examining changes in ribosomal densities after pairs, triplets, and so forth of rare codons, however, also indicates that consecutive rare codons do not systematically slow ribosomes (Figure S5 and Figure 3B). We achieve similar findings when defining rare codons according to their genomic frequency (Figure S2).

If the above results are correct, then we should also find that codon usage cannot explain ribosomal slowing when we compare sites within a given mRNA. Upon locating the highest and lowest ribosomal occupancy portions within a given mRNA, we determined whether the denser region was associated with a putative ribosome-slowing feature: lower tAI, or more rare codon pairs or rare 6-mers (two adjacent in-frame codons that, as a pair, come from the lowest 10% of all 6-mers within the genome) (see Methods, "The Relative Contributions of Charge, Folding, and Codon Usage to Extremes of Slowing Within Transcripts"). Considering all transcripts, the most slowly translated region within an mRNA in fact tends to be comprised of more optimal codons or fewer rare pairs, suggesting low codon optimality does not cause slowing (Table 1A,B and Figure S6). These results are

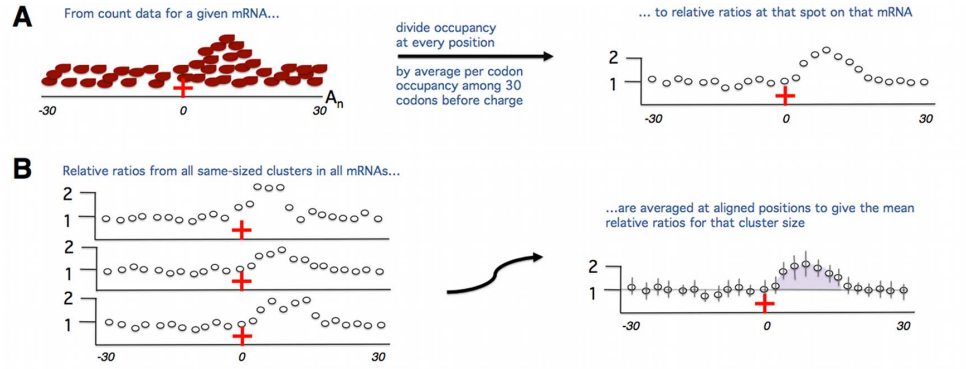

Figure 1. Visual overview of our plotting analyses. A feature of one codon encoding a positive charge as a potential slower of translation elongation is considered as an example. The feature of interest (here the encoded charge) must be surrounded by no other codons encoding positive charges for 30 codons in both directions so as to not interfere with our measurement of slowing due to the single encoded charge we have identified. (A) We start with footprint data, which we have stringently mapped to the codons surrounding the encoded positive charge of interest on the mRNA in which the encoded charge resides. We first count the ribosomal footprints mapping to each codon position in this area. We take the average of the ribosomal footprint counts among the 30 codons preceding (the start of) the feature. We consider the average footprint counts of these preceding 30 codons (r_{prec30}) to reflect the baseline speed at which ribosomes are translating before they reach the encoded charge. We then divide the ribosomal footprint counts in each of the 61 codon positions in this section of the mRNA by r_{prec30} to measure whether they are more densely or sparsely covered with ribosomal footprints in a given codon position relative to the density before the feature. Note the ratios prior to $x=0$ will tend to center around 1 as they will have been normalized by a value likely close to their own. We calculate these relative ratios separately for every feature cluster in every mRNA we identify as suitable for our analysis. (B) To ask whether there is a trend in slowing upon the translation of the feature of interest (the single positive charge in this example), we align all of the mRNAs with the feature of interest by (the start of) the feature. We determine the average relative change in ribosomal density upon translation of the feature by averaging each of the ratios calculated in (A) for each aligned codon surrounding the feature. It is these mean ratios we consider when we calculate the slowing effect (if any) of a given feature. The degree of slowing due to a feature is a function of both the magnitude of the footprint buildup on any one codon as well as the length along the mRNA that the buildup extends. We hence calculate the slowing due to the feature (here the single positive charge) by summing the area between the line $y=1$, which represents the baseline speed (see A) and the mean relative ratios between the start of the feature at $x=0$ and the point where the means cross $y=1$ again (highlighted purple area). If the line does not intersect with $y=1$ again by the end of the window ($x=30$), the entire area under the curve from $x=0$ to $x=30$ was used. We do not consider codons at $x>30$ as there may be positive charges encoded in this downstream region that we do not wish to interfere with our measurements. In some cases, not slowing but speeding will occur, indicated by ratios that are less than 1 (not shown). In this case, we calculate the degree of speeding similarly, by summing the area between the mean ratios and $y=1$.
doi:10.1371/journal.pbio.1001508.g001

not affected if we consider suboptimal codons to be those that are genomically infrequent (Table S1 and Figure S3). Nor do we find that transcript similarity to the yeast Kozak sequence can explain slowing within these regions (Figure S7 and Table S3). Additionally, as the difference in ribosomal occupancy between the two intra-transcript windows increases (and hence the presumed difference in the inferred ribosomal velocities between the two windows grows all the more), the already low proportion of transcripts for which tAI, genomic infrequency, or presence of rare pairs could possibly explain ribosomal pausing in fact decreases (Table 1A,B and Table S1). In other words, in transcripts that have the greatest differences in ribosomal densities along their length (as inferred from the highest and lowest ribosomal occupancy windows), and hence that contain the greatest degree of internal slowing relative to maximum translation speed, the most ribosomally occluded windows are even more likely to be comprised of more optimal codons. This indicates that not only is low codon optimality incapable of explaining ribosomal slowing in general, it is even less capable of explaining the greatest relative slowing within a transcript.

We note that the decrease in the ability of codon usage to explain slowing in the upper quantiles (Table 1A) is simply a side effect of differential amino acid usage between the two windows. When we control for differential amino acid content between the two windows, we no longer see the decrease in the ability of codon usage to explain slowing, but codon usage still remains unable to explain the slowing that is observed in any of the quantiles (Table S4). Thus, in addition to the above finding that codon usage becomes less able to explain slowing as the degree of slowing grows (as deduced from observed transcripts), this amino-acid-controlled analysis suggests that even if amino acid sequence had evolved in any other way, codon usage would still not be a factor in the slowing of ribosomes.

It is possible that codon usage could have different effects during different times of cell cycle if tRNA levels fluctuate [35]. We do not, however, detect a systematic influence of codon usage on ribosomal speed even under amino acid starvation conditions (Figures S8, S9, S10 and Table S5) when presumably tRNA charging levels are lower, making codon usage potentially more rate-limiting [36,37].

RNA Structure on Average Increases Ribosomal Occupancy Marginally

If neither codon usage nor consecutive rare codons can explain variation in ribosomal speed, then what can? As it has been suggested that transcript structure can impede ribosomes along the length of the transcript [28], we next investigated whether RNA structure might be the major contributor to slowing.

We used empirically determined (rather than computationally predicted) RNA structure data (PARS values, see Methods, "The Average Effect of Transcript Structure on Ribosomal Densities") [38]. *S. cerevisiae* protein-coding sequences were scanned for stretches whose average PARS value was 0 or negative (and hence tending to be single-stranded), which were immediately followed by a block of codons whose average PARS value was positive (i.e., with propensity for double-strandedness). The general contribution of folding to slowing was examined by

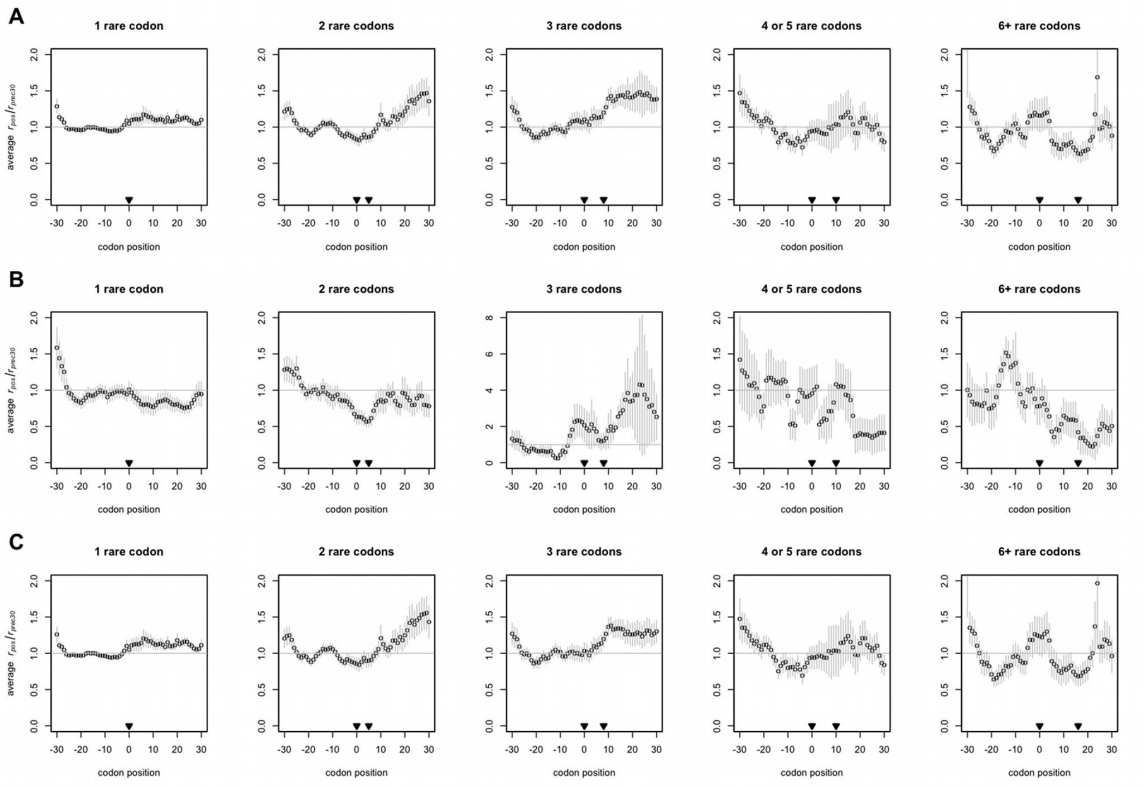

Figure 2. Clusters of rare codons do not tend to slow ribosomes. The first of the number of nonoptimal codons indicated always occurs at $x = 0$, and the rest, if any, may be found at points up to and including the codon indicated by the second arrowhead. The mean r_{pos}/r_{prec30}, or relative change in ribosomal occupancy, at each position across aligned transcripts ± s.e.m. is plotted. The horizontal at $y = 1$ represents the null expectation that positive charges do not alter ribosomal speed—that is, that ribosomes are, on average, as frequently present before the rare codon cluster as after it. The three-rare codon plot in (B) is plotted with different axes as it is an outlier. Some residual slowing is observed near $x = -30$ on all plots due to slowing elements (e.g., positive charges) that may be encoded just upstream ($x < -30$). (A) All genes with rare codon clusters. (B) Genes with rare codon clusters that have 0 or 1 positive charges coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density, with the bulk of the effect of positive charge removed. (C) Genes with rare codon clusters that have two or more positive charges in the last 30 codon positions plotted.
doi:10.1371/journal.pbio.1001508.g002

calculating the relative change in ribosomal density (r_{pos}/r_{prec30}) at each position of the identified region of a transcript, where r_{prec30} is the average ribosomal occupancy in the single-stranded block. We then take the average of this ratio across transcripts aligned by identified blocks of structure.

The method is similar to that used above with codons, but with one complication. In the case of codon usage, we have a prior expectation that any ribosomal pausing should occur while the ribosome is positioned over the "slow" codon. It is not immediately clear, however, where along the transcript we should expect any structure-induced pausing to take place. After translating an unstructured span of mRNA, will the ribosomal active site be able to get very close to the first double-stranded ribonucleotide it meets before it is finally slowed, or might pausing take place more 5′ if the ribosome progression is sterically occluded at some distance upstream? We investigated both hypotheses.

We cannot immediately distinguish between the possibilities that mRNA folding has an effect on ribosomal progression either upon or upstream of the folded ribonucleotides in question, as some degree of pausing is observed in both cases (Figure 4). But how strong is this slowing effect? Could mRNA folding account for the bulk of the variance in ribosomal speed observed along transcripts? We find, again comparing the slowest and fastest translated regions within a given mRNA, that not only is secondary structure incapable of systematically explaining the slowest regions of translation, but the presence of secondary structure decreases as the difference between the ribosomal density (i.e., difference in translation speed) of the two intra-transcript windows increases (Table 1C). Hence we conclude something other than mRNA folding must be responsible for the greatest slowing within transcripts.

Positively Charged Amino Acids Additively Slow Ribosomes on Endogenous Yeast Transcripts

We performed a parallel version of the codon cluster analysis to look for changes in ribosomal density after differently sized clusters of encoded positive charges (see Methods, "The Average Effect of Positive Charge on Ribosomal Densities"), calculating the average relative change in ribosomal density within a transcript (r_{pos}/r_{prec30}) after positively charged residues (lysine, arginine, or histidine) are

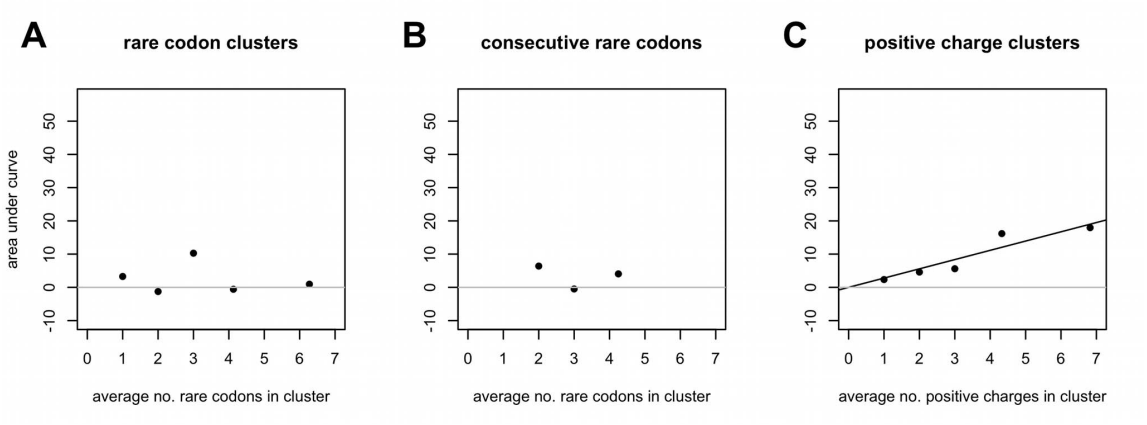

Figure 3. Positive charges show an additive (linear) trend in slowing ribosomes, but rare codons do not. The degree of slowing is a function of both the magnitude of ribosomal density and the length of transcript the slowing covers. Therefore, to measure any trend in the ability of either positive charges or codon clusters to slowing, the area between the curves depicting the average relative change in ribosomal density (r_{pos}/r_{prec30}) and the $y = 1$ null in Figure 2A, Figure S5A, and Figure 5, whether positive or negative, was summed between $x = 0$ (the beginning of the cluster) and the point where the plotted values intersect with $y = 1$ again, regardless of where the last charge in the cluster is (see Figure 1 for further explanation of the area under the curve). A positive value for the area under the curve indicates ribosomal slowing after the feature in question, while a negative value reflects faster movement. (A and B) Regression of *area under curve~size of cluster*, slope $p = 0.45$ and 0.33, respectively. (C) Regression of *area under curve~size of cluster* gives a slope of 2.81 ($p = 0.020$), $r^2 = 0.93$. To achieve such a regression slope in the set of genes used is significantly nonrandom ($p = 0.011$, Note S3).
doi:10.1371/journal.pbio.1001508.g003

added to a nascent peptide chain. The effect, note, should be a stalling after the codon specifying the charged amino acid as the stalling process is hypothesized to be an interaction between the charged amino acid and the charged exit tunnel [26,39].

We find that a single positive charge will slow the ribosome relative to the preceding sequence (Figure 5), regardless of whether the codon encoding the residue is A/G- or C-rich (Figure S11). Our findings show that at maximum (in real transcripts), ribosomes are more than twice as likely to be found at a given region of the transcript as before the addition of the cation to the polypeptide (Figure 5). The higher the density of positive charges in a peptide, the proportionally greater the effect (Figure 3C), in agreement with experimental findings that increasing the number of positive charges locally correspondingly increases ribosomal dwell time [26]. Our estimation of charge-induced pausing is conservative since some ribosomal density after charges is not included in the analysis if the mean ribosomal occupancy of the 30 codons preceding a charged cluster is 0 for a given transcript (our method in this case would require division by 0).

We can also test whether charge is responsible for slowing by noting that the pKa, and hence overall net charge, of histidine is lower than that of either arginine or lysine at physiological pH. Thus we should expect a weaker slowing effect due to histidine residues being added to the polypeptide. When we re-calculate the slowing effect after a single positive charge (as shown in Figure 5, first panel), but separate the single charges according to whether or not they are histidine, we indeed observe that histidine causes weaker slowing (Figure S12). The slowing effect after a single histidine residue, as calculated using the area under the curve method, is anywhere from 25%–78% (95% CI) of the slowing found after a single lysine or arginine. As histidine is used much less frequently than either of the other positively charged residues, we consider slowing after single positive charges to be the best comparator due to the larger sample sizes available. When we separate larger positive charge clusters according to their histidine content (at least one histidine in the two- or three-charge clusters, and at least two histidines in the four- or five-charge clusters), we note that the slowing due to the histidine-enriched group is always lesser than that after the histidine-free group (Figure S12).

If charge is a major determinant of ribosomal slowing, then it should be capable of explaining the regions of greatest translational pausing within transcripts (see Methods, "The Relative Contributions of Charge, Folding, and Codon Usage to Extremes of Slowing within Transcripts"). We find this is indeed the case. Of all the putative slowing features we consider, only positive charge is more often associated with the higher occupancy window within each transcript (Note S2). Breaking the comparisons into quantiles according to the magnitude of difference in ribosomal occupancy between each pair of windows further reveals that positive charge is the feature most often responsible for not just slowing when comparing between transcripts, but the greatest magnitude of slowing within any given mRNA. As the difference in ribosomal occupancy between the two windows increases, the window with the higher ribosomal occupancy tends increasingly to be the one with more positive charges (Table 1D). In fact the only clearly significantly overused amino acid in the higher occupancy windows is lysine, which is positively charged (Figure S13). This increase in ribosomal occupancy cannot be explained by physiochemical properties of other amino acids, namely hydropathy, negative charge, or polarity (Tables S6, S7, S8, S9). We note that even when both windows in a transcript have the same number of charges each, there is no predominant influence of tAI, rare codon pairs, or RNA structure on ribosomal slowing (Tables S10, S11, S12).

The Effect of Positive Charge Is Not Explained by Covariance with Codon Usage or mRNA Folding

The positive charge effects seen above could potentially be explained as covariate to codon usage bias, were, for example, codons specified by rare tRNAs especially abundant near those

Table 1. Only positive charge is systematically capable of explaining ribosomal slowing, including the severest slowing.

Score	Value	$q1_{\Delta r}$ (Count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 Test for Heterogeneity/*p* Value (Bonferroni Correction)
A. tAI score	1	590	597	563	525	0.13
	0	0	0	0	0	—
	−1	656	649	682	721	0.20
	Binomial test on +1 and −1 tAI score counts, *p* value (Bonferroni correction)	0.065 (0.26)	0.15	0.00082 (0.003)	3.1e-08 (1.2e-07)	—
B. rare pair score	1	175	179	144	86	3.0e-08 (8.9e-08)
rare 6-mer score		*127*	*106*	*86*	*45*	*9.5e-09 (2.9e-08)*
	0	858	885	905	1,037	0.00013 (3.8e-04)
		383	*403*	*424*	*503*	*0.00023 (0.00069)*
	−1	213	182	196	123	1.10e-05 (3.3e-05)
		199	*199*	*198*	*161*	*0.13*
	Binomial test on +1 and −1 rare pair score counts, *p* value (Bonferroni correction)	0.060	0.92	0.0056 (0.022)	0.013 (0.050)	—
		7.9e-05 (3.2e-04)	*1.1e-07 (4.4e-07)*	*2.5e-11 (1.0e-10)*	*<2.2e-16 (8.8e-16)*	—
C. PARS score	1	86	72	81	55	0.060
Conservative PARS score		*302*	*272*	*290*	*294*	*0.64*
	0	469	512	500	546	0.11
		0	*0*	*0*	*0*	—
	−1	154	124	127	108	0.036 (0.11)
		407	*436*	*418*	*415*	*0.78*
	Binomial test on +1 and −1 PARS score counts, *p* value (Bonferroni correction)	1.3e-05 (5.2e-05)	0.00025 (0.001)	0.0017 (0.0068)	4.0e-05 (0.00016)	—
		9.1e-05 (0.00036)	*7.6e-10 (3.0e-09)*	*1.7e-06 (6.8e-06)*	*6.3e-06 (2.5e-05)*	—
D. charge score	1	573	586	637	717	0.00014 (0.00043)
	0	258	259	236	207	0.0589 (0.18)
	−1	415	401	372	322	0.0038 (0.011)
	Binomial test on +1 and −1 charge score counts, *p* value (Bonferroni correction)	5.6e-07 (2.2e-06)	4.3e-09 (1.7e-08)	<2.2e-16 (8.8e-16)	<2.2e-16 (8.8e-16)	—

Quantiles of the difference in average ribosomal density between the most highly occupied and most lowly occupied windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. Related yet alternative ways of calculating both the rare pair and PARS scores are given in italics (see Methods, "The Relative Contributions of Charge, Folding, and Codon Usage to Extremes of Slowing Within Transcripts" for details). A low codon optimality, if anything, tends to pair more with the less dense (faster translated) window. Similarly, not only do rare pairs and rare 6 -mers tend to be found more often in the faster translated window, but their presence decreases as the difference in degree of ribosomal slowing grows. Additionally, a greater likelihood of transcript secondary structure at or just before the identified window is associated not with the more occluded windows, but with the less dense (faster translated) ones, and the presence of secondary structure in fact decreases as the difference in ribosomal slowing between the windows increases. Positive charge, however, is consistently associated with the higher density (more slowly translated) window, and increasingly so as the difference in densities between the two windows becomes larger. Window pairs that have the same number of charges each (charge score, 0) do not show such a trend between quantiles.

doi:10.1371/journal.pbio.1001508.t001

specifying positively charged residues. Given the absence of evidence for codon usage bias to affect translation rates, this now seems unlikely. To nonetheless test whether this is the case, we examined patterns of codon usage in the vicinity of positive charges similarly to the manner in which we investigated changes in ribosomal occupancy after positively charged clusters above. Thus if nonoptimal codon usage were causing the slowing patterns after encoded positive charges observed in Figure 1, we should see, on average, a relative decrease in tAI in those sites with elevated ribosomal occupancy. Contrary to this expectation, however, the trend for ribosomal occupancy to increase after positive charges (Figure 5) is independent of patterns of codon usage (Figure S14).

It is also possible the slowing effects observed after positive charge clusters in Figure 5 occur ancillary to mRNA secondary structure, as such structure may have some slowing effect (Figure 4). Again we allow for mRNA folding to impede the flow of ribosomes starting either locally or 10 codons upstream (in the case that local double-strandedness creates a structure within the

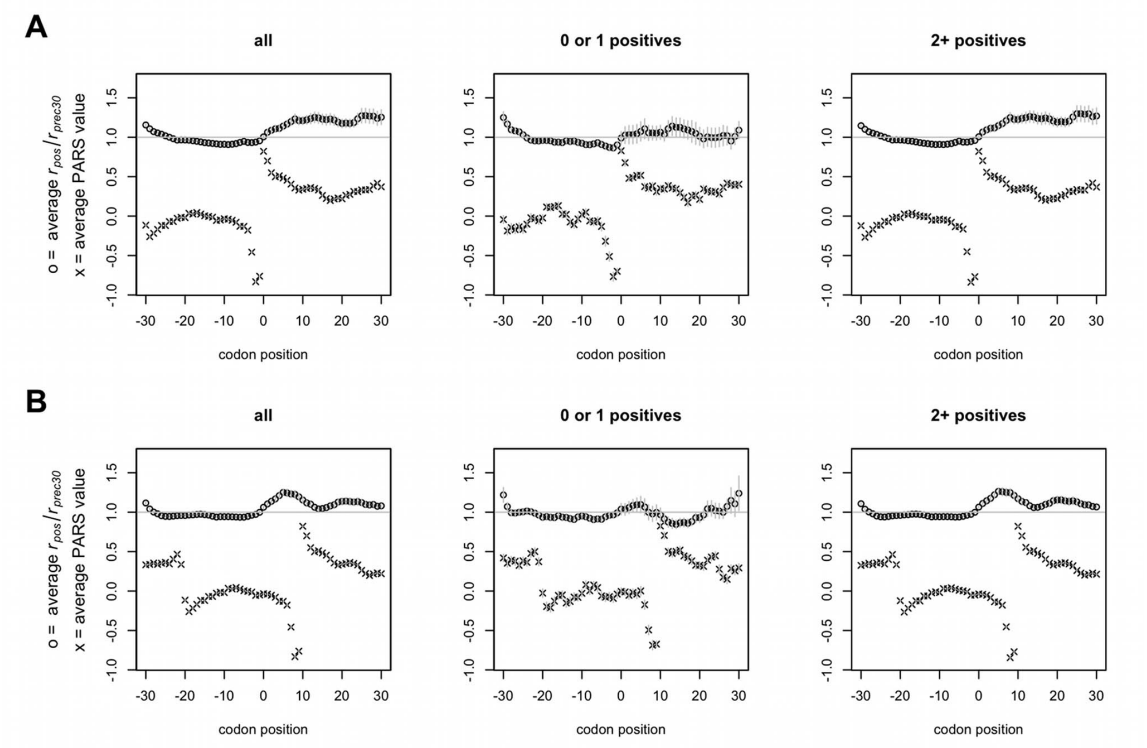

Figure 4. Ribosomes travelling along single-stranded RNA are not greatly retarded upon traversal into double-stranded structures. PARS values >0 denote structured mRNA, <0 single-stranded. All averages plotted (± s.e.m.) are calculated across transcripts aligned by blocks of mRNA structure. The slowing of ribosomes ($r_{pos}/r_{prec30}>1$) relative to the preceding 30 codons starting from both the beginning of double-stranded structure (A) and 10 codons upstream of the same regions of double-stranded structure (B) are shown. In both cases, there is a degree of translational pausing observed upon the transition into folded mRNA, although some of this slowing may be caused by the presence of two or more positive charges encoded in the folded area ($0 \geq x \leq 30$).
doi:10.1371/journal.pbio.1001508.g004

transcript that sterically occludes ribosomes from progressing further toward codons within the folded structure). We find that patterns of transcript secondary structure near positive charge clusters are unable to explain the pausing after translation of positive charges (Figure S14). Hence we argue that mRNA folding cannot explain the slowing seen in Figure 5, which is

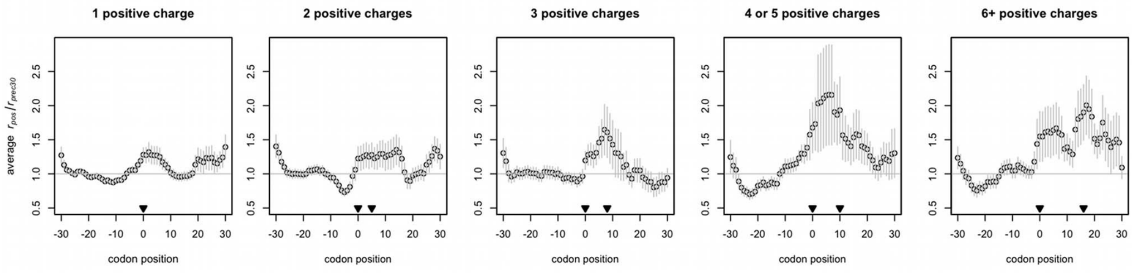

Figure 5. Positive charges slow ribosomes. The first of the positive charges indicated always occurs at $x=0$, and the rest, if any, may be found at points up to and including the codon indicated by the second arrowhead. r_{pos}/r_{prec30} is the ribosomal occupancy at position x normalized by the average occupancy of the 30 codons preceding the encoded positively charged cluster within the same transcript. The mean r_{pos}/r_{prec30}, or average relative change in ribosomal occupancy, at each position across aligned transcripts ± s.e.m. is plotted. The horizontal at $y=1$ represents the null expectation that positive charges do not alter ribosomal speed; in other words, that ribosomes which translate in positive-charge free peptides are, on the average, as frequently present before the charge cluster as after it.
doi:10.1371/journal.pbio.1001508.g005

perhaps not surprising given its apparently weak effect on the whole (Figure 4).

Covariance with Positive Charge Does Explain Some of the Slowing Observed After RNA-Level Features

Given that positive charge slows ribosomes, we should expect that some of the (relatively weaker and/or inconsistent) ribosomal slowing at rare codon clusters or transcript secondary structure might in fact be due to the presence of uncontrolled-for positive charge. We find this to be the case. When groups of rare codons that are followed by either a lesser or greater number of positive charges are plotted separately, it is clear that rare codon clusters do not in and of themselves slow ribosomes (Figure 2B) but that the apparent (yet unsystematic) slowing in Figure 2A is in fact due to the presence of positive charge after some of the codon clusters (Figure 2C). Similarly, sorting by the number of positive charges present after a cluster reveals that some of the slowing observed at structured regions of transcript is likely due to previously unaccounted-for positive charge (Figure 4).

Discussion

We find that codon usage and transcript secondary structure do not substantially affect ribosomal velocities systematically across endogenously occurring transcripts. Although it has been suggested that amino acid starvation might increase the ability of codon usage to modulate ribosomal speed [37], we find no such effect upon examination of ribosomal footprints taken from amino-acid-starved yeast (Figures S8, S9, S10 and Table S5). We do not, however, wish to assert that codon usage and RNA structure can never affect translation rates. Certain secondary structure configurations may substantially impact ribosomal flow. As regards codon usage, if we return to the original logic by which codon usage was thought to affect translation rates, we can both see where the prior logic was misleading and in turn can predict when codon usage should slow ribosomes.

The classical logic supposes that because common codons are specified by abundant tRNAs, the waiting time for the ribosome to capture the necessary tRNA must be lower for "optimal" or common codons. The key parameter, however, to determine waiting time is not the absolute tRNA abundance (as often considered) but the tRNA availability. We note, similarly to Qian et al. [16], that if codons are used in proportion to tRNA availability [40], then this could dampen any pausing effect, since rare codons matching rare tRNAs will not be as rate-limiting as if they were used more often. Put differently, if highly abundant transcripts all require the same tRNA, then this acts as a drain on the availability of that tRNA. This can be described in terms of supply and demand economics. In the case of rare codons in lowly expressed transcripts, the supply (the pool of tRNA) is small and the demand (number of codons requiring that tRNA at any given time) low. For a common codon in an abundant transcript, the supply (tRNA pool) is large but the demand is also large.

We can then imagine an equilibrium situation in which the ribosome waiting time is the same for all codons as the demand (absolute codon abundance in transcripts) and supply of tRNAs are balanced. This is consistent with our observation that, under normal growth conditions, codon usage does not predict ribosome occupancy. However, the same model can predict that under abnormal conditions, we might see an effect as the situation has been forced far out of supply–demand equilibrium. Greatly overexpressing a transcript rich in rarely used codons should slow the ribosome as the demand for the rare tRNAs now exceeds supply. Likewise, we expect that gross modification of tRNA pools should have gross effects on translational speed as the system has been shifted away from the demand–supply equilibrium. This distinction between normal (equilibrium) and experimentally forced (nonequilibrium) conditions makes good sense of the prior literature, where reports of an effect of codon usage on translational velocity involved experimentally forced conditions (for review, see Note S1).

Further evidence that the impact of codon/tRNA abundance is buffered comes from the report that some codons whose aminoacyl-tRNAs are selected either intrinsically rapidly or slowly by the ribosome have either low or high tRNA concentrations within the cell, respectively [41], suggesting that intrinsic differences in the translation speeds of certain codons are not accentuated but rather compensated for. The evidence for codon usage/tRNA buffering indirectly suggests either that some property other than speed causes selection on codon usage (e.g., accuracy of translation [18–21]) or that selection for speed occurs when the demand–supply balance is perturbed, for example when selection acts on growth rates and favor duplications of tRNAs. That codon usage also has little or no effect on ribosome velocity in mammals [15] as well as yeast is then, in retrospect, perhaps not so unexpected.

Our results are consistent with the interaction of the cations in the protein with the ribosomal exit tunnel [25,26], a model supported by the stalling being displaced from the location on the mRNA of the codons specifying the positive charge. Our results also indicate that positive charge, more than other chemical or biophysical properties of amino acids (see Tables S6, S7, S8, S9), is key. While some highly conserved amino acid sequences have been shown to interact with the ribosomal tunnel to stall translation in order to regulate the specific gene product they control (see, e.g., [42–44]), our results suggest a fundamental feature of proteins that slows ribosomes regardless of sequence context (either the local amino acid sequence or the gene in which they reside) and without the addition of trans acting factors.

A general slowing of translation due to positive charge has ramifications for the evolution of the poly-A tail. If translated, the poly-A tails results in a long run of positively charged lysines. This is expected to stall run-on ribosomes [39]. This stalling may glue the aberrantly translated peptide to the ribosome, preventing potentially toxic products from diffusing into the cell and/or permit tagging of the peptide in the nascent chain–ribosome complex with a signal for degradation, as observed [26,39].

Our results are consistent with translation of poly-A tails stalling ribosomes. Extrapolating the linear trend for larger clusters of positive charges to additively slow ribosomes (reported in Figure 3C), we note that a poly-A tail of 80 consecutive adenines (~27 lysines) in yeast [45] should slow translation at least 4-fold more than that observed in clusters of six or more positive charges (Figure 3C), probably halting it. This is in line with experimental work showing that while nonstop mRNAs without poly-A tails are efficiently translated [46], translation of polyadenylated mRNAs lacking stop codons or full 3′UTRs is repressed after initiation [47]. Similarly, inserting a poly-A tract into a coding sequence represses translation post-initiation, but not on account of rapid mRNA decay [39]; a similar finding was reported for 3′ poly-A tails [48]. Recently, it was shown that translation of 12 consecutive basic amino acids inserted into a reporter gene causes not only translation arrest but degradation of the polypeptide [49].

Why is the tail poly-lysine if any positive charge will do? The reason is likely to be found at the DNA sequence level. Of all codons encoding positive charges, only lysine possesses a codon that is a triplet repeat of a single nucleotide (AAA) and therefore may be added simply and sequentially by a single enzyme.

Moreover, the triplet repeats form a homogenous run of adenines, meaning that positive charges will still be added to the nascent chain (and hence stall ribosomes) no matter how the stop codon is missed, be it by failure to interpret the stop when in-frame or owing to frame-shifting. This may have less relevance in species with long 3′UTRs, in which an alternative stop may be found with the UTR, but in the ancestor in which the poly-A tail evolved, if 3′UTRs were short, then this sandtrap for ribosomes may have been of considerable benefit.

It is noteworthy that bacteria, which for the most part lack poly-A tails, have an alternative mechanism (tmRNA) to tag and destroy proteins resulting from frameshifting or stop codon readthrough [50]. Stalling initiated by positive charges resulting from translation of poly-A tails in eukaryotes and tmRNA system in prokaryotes may be functionally equivalent modes of error correction [51].

Methods

Ribosomal Density Data

Both sequenced ribosomally protected fragments and sequenced fragmented total mRNA for *S. cerevisiae* dataset GSE13750 [29] were downloaded from the NCBI Gene Expression Omnibus at www.ncbi.nlm.nih.gov/projects/geo. The rich media and amino-acid-starved sets were considered separately. Annotations of the *S. cerevisiae* S288C genome as available on June 22, 2008 (the build used by Ingolia et al. [29]) were obtained from the *Saccharomyces cerevisiae* Genome Database (www.yeastgenome.org). Only protein-coding sequences of nondubious classification were considered, giving 6,262 genes for potential analysis. Any sequences containing nonsense codons or that were not multiples of three were excluded. The sequences were further filtered to only allow the standard or alternative start codons indicated in NCBI genetic code Table 1 from http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter = tgencodes, leaving 6,215 sequences for analysis. The chromosomal location and coordinates of the sequenced fragments given in the original dataset were used in combination with the start and stop coordinates of genes from the annotations to determine which fragments map to which genes, and in the case of the footprint fragments, where along the coding sequence the protected area lies.

Since the probability of sequencing error in a stretch of ~28 nucleotides is quite low (for runs <50 bp on Genome Analyzer 2, error rates are expected to be around 1%), only one mismatch between the sequenced fragment and reference genome sequence was allowed. All fragment counts were taken as the average value of the two experimental replicates. Fragments that were sequenced at least once in one replicate but not listed in the other were marked as having an expression count of 0 at analogous positions in the latter replicate. In the case of fragments that map to more than one possible genomic location, it is impossible to tell which are the true areas covered by ribosomes. In order to avoid the introduction of false-positive ribosomal occupancies en masse into the dataset, which could systematically bias the types of sequences that are occluded, only footprints that mapped uniquely to one location in the reference genome were considered.

In line with Ingolia et al. [29], we assigned footprints to protein-coding genes of nondubious classification if the first base of the footprint mapped to 16 nt before the first base or 14 nt before the last base of the gene, in order to take account of which area of the footprint is likely in the ribosomal active site. Since the chance of sequencing another fragment from a stretch of coding sequence increases as a function of gene length, mRNA fragment counts were normalized by dividing by gene length for the relevant gene. In addition, the footprint counts were then divided by the normalized mRNA counts mapping to that gene to obtain per-transcript ribosomal densities (indexed by location along the transcript). We performed this normalization by mRNA to ensure that differences in occupancies we calculate (see Methods, "The Average Effect of Positive Charge on Ribosomal Densities") are not an artefact of mRNA levels. This left us with a final 5,430 filtered genes with footprint coverage mapped per codon pair per transcript.

The Statistical Approach to Handling the Occupancy Data

We note that there are two previous studies [27,28] that examined this ribosomal footprint data [29] and found a role for codon usage in modulating ribosomal speeds, mainly by detecting a correlation between the local codon optimality along a transcript and the corresponding local density of ribosomal footprints. We cannot offer a reason why these studies produce such a finding, namely because they do not detail their methodology concerning the ribosomal footprint data, including whether they used all or just a subset of all the sequenced footprints (e.g., dependent on footprint length, the number of mismatches to the genome reference sequence allowed, or the number of places in the genome to which a single footprint could simultaneously map). Another study [16] that examined the same data contradicted the finding that codon usage affects ribosome velocity, highlighting the importance of methodology in the analysis of ribosomal profiling data. This opposing study [16], however, may have mapped footprints to multiple genomic locations and also considered only footprints 28 nt in length in an attempt to precisely map which codon is in the A-site and hence selecting an aminoacylated tRNA. We are not confident that the interpretation of ribosomal footprint data allows for such specificity, as the interpretation of where the A- and P-sites are along the ribosomal footprint seems to be inferred from the average footprint length obtained during initiation and termination [29], whereas the conformation of the ribosome and hence footprint obtained during elongation may differ. Also, it has since been noted that elongation inhibitors, such as cyclohexamide, which was used in the creation of the dataset under consideration [29], alter the conformation of the ribosome, leading to advised caution in determining position-specificity from individual footprints [15]. For these reasons, we consider it optimal to stringently map footprints to a single location in the genome, thus preventing the introduction of a false correlation between certain codons and ribosomal density, and to consider all of the sequence that is occluded by the footprint instead of attempting to pinpoint the location of a structural site in the ribosome from the artefact of the footprint.

To determine what determines occupancy, we could consider some general linear model in which we employ multiple parameters (local codon usage, local RNA stability, and local charge density) to predict occupancy on a codon-by-codon basis. However, such models assume that the data points are independent. Owing to the nature of the data (ribosomes sit over spans of sequence), the occupancy seen at one codon by necessity is nonindependent of that seen at neighboring codons. Thus, such methods are not generally valid. To overcome the nonindependence problem, we do not consider each codon as a separate data point. Rather we consider the dimensions of the spans of increased relative occupancy and consider how trends in the dimensions of these spans correlate with the density of the potentially slowing feature in question, an approach that we outline in Figure 1 and below.

We consider for any feature (e.g., a cluster of codons or positive charges) the start position of this feature (position $x = 0$). We then define for each codon at and after the start of the feature ($x \geq 0$) how the occupancy is related to the mean occupancy of the 30 codons upstream of $x = 0$ within the same mRNA as the feature. We define the relative occupancy of any given codon (r_{pos}/r_{prec30}) as its occupancy (r_{pos}) divided by the mean occupancy of the 30 codons prior to the considered feature (r_{prec30}). For plotting purposes, we also normalize the occupancy of all codons 5′ of the focal position at $x = 0$ by the same r_{prec30} value. Dividing by pre-cluster ribosomal densities to obtain a ratio normalizes for differences between transcripts such as expression level, accommodates and normalizes for differences in ribosomal density that may be caused by characteristics of upstream sequence, and allows for comparisons of the relative change in ribosomal movement across different mRNAs. These relative occupancy ratios, which we calculate surrounding every identified feature, thus represent the speeding (if the ratio is <1) or slowing (if the ratio is >1) of ribosomes after a given feature as they translate that portion of that gene. After calculating the relative ribosomal occupancy ratios surrounding a feature within all available mRNAs, we then align these mRNAs by the start of that feature and calculate the average relative occupancy ratios across transcripts. We plot these averages such that y at codon $x = 1$ is the mean ratio of all the observations across multiple RNAs at $x = 1$, codon $x = 2$ is the mean ratio across multiple RNAs at $x = 2$, and so on. These plots then present a span of increased relative occupancy or of decreased relative occupancy following the start of the feature at $x = 0$ across all instances of that feature available to our analysis.

The Average Effect of Codon Usage on Ribosomal Densities

As tRNA gene copy number has been shown to strongly correlate with tRNA abundance [27,52], preferential use of codons that base-pair to the anticodons of high-copy tRNAs is taken to reflect adaptation of coding sequence to the tRNA pool and hence optimal codon usage for translational efficiency and/or accuracy. Each codon can then be ascribed an adaptiveness value (W_i) [32]. The tRNA adaptation index (tAI) is the geometric mean of the scores for the constituent codons and is then a measure of the degree to which protein-coding genes use codons corresponding to tRNA isoacceptors with high gene copy numbers within a given genome (although, see also Figures S1, S2, S3 and Table S1 for analyses of rare codons where "rare" is defined as genomically infrequent) [32]. The codonR package to calculate tAI was downloaded from http://people.cryst.bbk.ac.uk/~fdosr01/tAI/index.html on May 7, 2011. Yeast tRNA genes were obtained from the UCSC Table Browser [53] at http://genome.ucsc.edu/cgi-bin/hgTables. Statistical calculations for the tAI (and for other analyses generally) were done in R [54].

As a gene with just one codon would have a tAI value equal to W_i of that codon, we refer in the text to a codon's tAI value. In the main text we define rare codons to be those in the lowest quartile of W_i values as derived for yeast (CGA, ATA, CTT, CTG, CTC, CGG, AGT, CCC, GCG, AGC, CCT, TCG, TGT, ACG, and GTG). We interchangeably use "rare" for "non-optimal" as the frequency of codon usage in yeast is roughly proportional to the numbers of tRNAs that can decode them [40]. Protein-coding sequences were scanned for single rare codons; two rare codons anywhere within a five-codon stretch, three rare codons within eight codons, four or five within 10, and six or more within 16. The cluster specifications outlined here were chosen to maximize cluster sample sizes while incorporating the following caveat: we required that a block of 30 non-rare codons had to precede the identified codon clusters and that no other rare codons could be present in the next 30 codons apart from those in the identified cluster. When investigating consecutive rare codon clusters in a parallel analysis, we required that no consecutive rare codons could be present in the surrounding ~60 codons apart from those in the identified cluster (single rare codons were permitted as otherwise sample sizes would be far too small). As noted above, the first rare codon in the cluster is always considered to be at position $x = 0$.

As there are not enough rare codon clusters that are isolated from the ribosome-slowing effects of positive charges, we were unable to introduce the requirement that no positive charges be present in the vicinity of the rare codon cluster. Instead, we split the rare codon clusters into two groups—those that had two or more positive charges coded for in the sequence following the rare cluster, and those that had either zero or one positive charge—and plotted the results for these groups separately.

As noted above, we perform a normalization with respect to the local ribosomal occupancy. Within a given mRNA, the relative increase or decrease in ribosomal density (r_{pos}/r_{prec30}) at each position surrounding a rare cluster was calculated by dividing the measured ribosomal density at each codon position (r_{pos}) by the average ribosomal occupancy of the thirty codons preceding the first rare codon in the cluster (at position $x = 0$) within that same mRNA (r_{prec30}). The average relative change in ribosomal occupancy (mean r_{pos}/r_{prec30}) at a given position during/after a cluster was then calculated by aligning all identified regions of a given cluster size according to the first codon present in each cluster and calculating the average ratio (i.e., increase or decrease in measured ribosomal occupancy) in positions increasingly distant from the aligned clusters (a schematic of this approach is contained in Figure 1).

The Average Effect of Transcript Structure on Ribosomal Densities

We used experimentally and not computationally determined RNA structure data. By exposing transcripts independently to endonucleases specific for single- and double-stranded RNA, the degree to which individual nucleotides of an mRNA are involved in intramolecular secondary structure has been experimentally quantified [38]. The resulting metric is "parallel analysis of RNA structure" (PARS) values, with higher values (positive) indicating a propensity for secondary structure and lower (negative) values signifying lack thereof. The authors show the PARS metrics along transcripts in yeast globally correlate with the degree of single- versus double-strandedness predicted by the Vienna Package. PARS values for yeast transcripts were downloaded at http://genie.weizmann.ac.il/pubs/PARS10/pars10_catalogs.html. PARS values corresponding to CDS regions were determined relative to the local coordinate file from the same website.

S. cerevisiae protein-coding sequences were scanned for stretches 30 codons in length whose average PARS value was 0 or negative (and hence tending to be single-stranded), which were immediately followed by a block 31 codons in length whose average PARS value was positive (i.e., with propensity for double-strandedness). To ensure a clear transition from single- to double-stranded structure upon averaging across transcripts, we added the requirement that the last codon in the first 30-block have a negative PARS value and that the first codon in the subsequent 31-block have a positive PARS value. Only nonoverlapping blocks (61 codons in length) were retained, with priority given to those with the highest combined number of negative PARS value in the first 30 codons and positive PARS value in the latter 31 codons.

The general contribution of folding to slowing was then examined by calculating r_{pos}/r_{prec30} (as described above) at each position and then taking the average across aligned single-stranded into double-stranded blocks. In the first instance of such a test, we investigated the hypothesis that the ribosome closely approaches the base of the double-stranded structure such that the ribosome is positioned closely over the first double-stranded ribonucleotide (i.e., at the beginning of the 31st codon out of 61) by the time slowing occurs. Here the first 30 codons in the identified block are classed as the preceding 30 codons before slowing might occur. We then repeated the analysis examining whether pausing of the ribosome might occur somewhat further upstream—for example, if the mass of the ribosome sterically hinders it from progressing at its normal rate even before the double-stranded ribonucleotide approaches the active site. In this second analysis, we used the same identified blocks as above, but moved the potential point of slowing to 10 codons upstream of the first codon with a positive PARS score. Hence the preceding 30 codons used in this case to normalize nearby ribosomal densities were also shifted 10 codons upstream as well.

The Average Effect of Positive Charge on Ribosomal Densities

Changes in rates of translation were measured by calculating the relative change in ribosomal densities that occurs within a transcript, on average, after positively charged residues (lysine, arginine, or histidine) are added to the nascent peptide chain. Such an effect should be observed at or after the encoded charge(s) in the mRNA as the positively charged amino acid travels down the exit tunnel. To test for an additive effect of charge on ribosomal density, *S. cerevisiae* protein-coding sequences were scanned for single positively charged amino acids, two positively charged residues anywhere within five amino acids, three positively charged residues within eight amino acids, four or five positively charged amino acids within 10 amino acids, and six or more positive charges within 16 amino acids, with the first positively charged residue always considered to be at $x = 0$. As in the case of the rare codon cluster analyses, these loosely defined cluster specifications were chosen to maximize the sample sizes available of clusters containing different numbers of positive charges.

To eliminate interference from charged amino acids outside these charged clusters, we required that a block of 30 non-positively charged amino acids precede the identified positive-charge clusters, and that no other positively charged amino acids be present in the next 30 amino acids apart from those in the identified cluster. Thirty residues were chosen as this is approximately the length of extended peptide that the ribosomal exit tunnel can accommodate [55,56]. Thirty non-basic residues therefore should provide a baseline ribosomal occupancy reading, and hence inference of the speed of translation, before the positively charged residues are added to the peptide chain and enter the exit tunnel.

The relative increase or decrease in ribosomal density at each position (r_{pos}/r_{prec30}) was calculated for each transcript with an encoded positive-charge cluster. The average relative change in ribosomal occupancy (mean r_{pos}/r_{prec30}) at a given position during/after a cluster was then calculated across regions aligned by similar-sized clusters (see also Figure 1 for a visual of this approach). Regarding our methodology, we find that noise in footprint density is not a problem for our analysis as we see similar findings when we consider genes with either low or high footprint coverage (Figure S15).

The Relative Contributions of Charge, Folding, and Codon Usage to Extremes of Slowing Within Transcripts

The above methods start by locating the appropriate putative ribosome-slowing feature within transcripts and then measures changes in ribosomal occupancy surrounding them. A complementary approach is to look to large changes in ribosomal density and then ask whether positive charges or rare codons are more often associated with the denser, putatively more slowly translated regions. Such an approach is best carried out on a within-mRNA level, as this normalizes for differences in overall expression levels across genes. Within each gene for which we retained ribosomal protection data (see Methods, "Ribosomal Density Data"), we located the two nonoverlapping 10-codon windows (approximately the length of RNA a ribosome footprint spans [29]) with the highest and lowest average ribosomal occupancy in that transcript. To circumvent the arbitrariness of choosing the location of the low-occupancy 10-codon window in a transcript for which there may be multiple possible windows with no footprint data available (i.e., a footprint count of 0), we added the requirement that experimental protection data exist for each codon in the window.

For each window in the pair, we recorded the average ribosomal occupancy as well as (1) the tAI; (2) the number of adjacent, nonoverlapping pairs of rare codons; (3) the number of positive charges encoded within and up to five codons upstream of the window (since a charge added while the ribosome was a few codons upstream should still be present within the exit tunnel); (4) the number of rare 6-mers (defined to be the lowest 10% of all possible in-frame 6-nt sequences within open reading frames); and (5) propensity for transcript secondary structure. We included 6-mers here, as while individual codons may be rare, it does not necessarily follow that two adjacent rare codons are just as rare of a combination, and thus examining the contribution of rare 6-mers provides an extra level of stringency in assessing the role of codon usage. As secondary structure either at the codon in question or downstream of the codon in question might pause ribosomes (see Results), we considered the PARS values not only within but also for an additional 10 codons downstream of each originally identified 10-codon window.

If codon usage bias is modulating ribosomal speed, we expect to observe the main effect over and locally surrounding the codon in question, whereas we expect, if charge indeed is influencing ribosomal velocity, to observe a downstream effect of positive charge on ribosomal density as the cation travels further down the negatively charged exit tunnel. Thus we also counted any positive charges encoded in the five codons immediately preceding the identified windows since a charge added while the ribosome was a few codons upstream should still be present within the exit tunnel. Conversely, as secondary structure either at the codon in question or downstream of the codon in question might pause ribosomes (see Results), we considered the PARS values not only within but also for an additional 10 codons downstream of each originally identified 10-codon window. Since the interpretation of PARS values may be somewhat more labile (as not only the magnitude but the sign of the values may have meaning), we tested whether double-stranded structure might associate with the more dense window in two different ways. We used two methods of measuring propensity for transcript structure. In the first method ("PARS score"), an average PARS value for a window ≤ 0 means the window is single-stranded, and an average PARS value >0 means the window is double-stranded; the exact magnitude of the PARS value is disregarded beyond this. In the second method ("conservative PARS score"), smaller changes in the magnitude of PARS values count more: the mean PARS value is calculated

for each extended (20-codon) window, and the means are then compared.

The ability of each metric to explain the difference in average ribosomal occupancies between the two windows was then assessed by asking how often the window with more of a potentially ribosome-slowing feature was also the window with greater occupancy. For example, if the window with the higher ribosomal occupancy paired with the less optimal (lower) tAI—which would be expected if less optimal codons do in fact slow ribosomes—then the gene was assigned a tAI score of 1; if the higher occupancy window paired with the more optimal tAI, indicating tAI is not a good predictor of increased occupancy, the a tAI score of −1 was assigned; and if the tAI was the same in the two windows, a score of 0 was given. Similar tests were performed independently on the number of rare codon pairs, PARS metrics, and number of positive charges associated with each window, with more rare pairs/more positive charges in the more occupied window—and hence potentially capable of explaining the elevated ribosomal density—each being scored 1, fewer being scored −1, and the same number in each window scored 0.

There are two potential complications of this method that we are able to address and dismiss. Firstly, although there is a tendency for ribosomal occupancy to decrease, on average, along the length of transcript [29], the correlations we report hold when we test only transcripts in which the high ribosomal occupancy windows are downstream of the low ribosomal occupancy windows (higher occupancy and increased positive charge, Spearman rho 0.12, $p = 0.00031$; higher occupancy and an excess of rare pairs, Spearman $p = 0.62$; an excess of rare 6-mers, rho = −0.06, $p = 0.07$; higher occupancy and lower tAI, Spearman rho −0.15, $p = 2.4e\text{-}05$). This, along with the additive pattern in Figure 3C, shows that the correlation between positive charge and increased ribosomal density is not methodological artefact.

Secondly, a window might have an apparently low average ribosomal occupancy if in fact there were ribosomal footprints that should have been assigned to that region of the transcript, but which was ultimately excluded from the analysis if the footprint mapped to multiple genomic locations. To this we note that the same analysis, when redone allowing the low-occupancy window to have a footprint count of zero, still gives similar results (Table S13). To further test this possibility, we created a list of nonredundant locations. These are sites in each transcript for which all mapping footprints were uniquely mapping to that location. In other words, no footprint data were excluded from being mapped to these sites because it also mapped somewhere else in the genome. Redoing the window comparisons analysis using the nonredundant locations, we find the results (Figures S16, S17 and Table S14) qualitatively match our original results using the dataset described above (see Methods, "Ribosomal Density Data"). Hence, we consider that our method fairly infers the contribution of different sequence features to ribosomal slowing.

Supporting Information

Figure S1 Figure 2 redone using rare codons defined according to genomic frequency shows rare codons do not slow ribosomes. In the main text, we investigate whether nonoptimal codons—that is, those with low tAI scores—might slow codons and find that they do not. To ensure that our finding that these "rare" codons do not slow ribosomes does not simply hinge on our definition of "rare," we have repeated the analysis using an alternative definition. Here, we define "rare" codons according to their actual frequency in the genome as measured from our set of filtered genes. This rare set, of equal size to the rare tAI set, comprises the following codons: CGG, CGC, CGA, TGC, CCG, CTC, GGG, GCG, CGT, CCC, CAC, TGT, ACG, TCG, and AGG. We find that rare codons, where rare means genomically rare, do not slow ribosomes when in clusters (single rare codons; two rare codons anywhere within a five-codon stretch; three rare codons within eight codons; four or five within 10; and six or more within 16). Note slowing should be observed over, not after, the rare codon(s). (A) All genes with rare codon clusters. Regression of *area under curve~number of rare codons in cluster*, slope = −0.79, $p = 0.080$. Regressions were performed as detailed in the main text (see Figure 1 for a description of the calculation of area under the curve). We note even if p were significant, the slope would be negative, whereas if rare codons did slow ribosomes, we should expect to see a positive slope. (B) Genes with rare codon clusters that have 0 or 1 positive charge coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density, with the bulk of the effect of positive charge removed. (C) Genes with rare codon clusters that have two or more positive charges in the last 30 codon positions plotted.
(PDF)

Figure S2 Consecutive rare codons, where rare is with reference to genomic frequency, do not slow ribosomes. In the main text, we investigate whether nonoptimal codons—that is, those with low tAI scores—might slow codons and find that they do not. To ensure that our finding that these "rare" codons do not slow ribosomes does not simply hinge on our definition of "rare," we have repeated the analysis using an alternative definition. Here, we define "rare" codons according to their actual frequency in the genome as measured from our set of filtered genes. This rare set, of equal size to the rare tAI set, comprises the following codons: CGG, CGC, CGA, TGC, CCG, CTC, GGG, GCG, CGT, CCC, CAC, TGT, ACG, TCG, and AGG. The consecutive rare codons in considered codons are present between the first and second arrowheads. See Figure 1 for a description of the calculation of the area under the curve. (A) All genes with rare codon clusters. Regression of *area under curve~number of rare codons in cluster*, slope = −9.8, $p = 0.32$. (B) Genes with rare codon clusters that have 0 or 1 positive charge coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density, with the bulk of the effect of positive charge removed. (C) Genes with rare codon clusters that have two or more positive charges in the last 30 plotted codon positions.
(PDF)

Figure S3 Codons that are overused in high-ribosomal occupancy windows are not "rare" according to genomic frequency. In some supplemental analyses, we examine whether "rare" codons slow ribosomes and define "rare" as the quartile of those most infrequent codons in the genome. To ensure there is not a problem with this definition, we have examined the difference in trends of codon usage at large between the two windows. (A) Tallies of all the codons used among the high-occupancy and low-occupancy windows within each gene (including the preceding five codons before each window) were kept separately. We plotted the counts for each codon in the high ribosomal occupancy window versus the counts in the low occupancy window and have color-coded the codons according to their frequency (see also Figure S6 for rare codons defined according to their tAI). If all codons are used equally among the slowly translated and quickly translated windows, then the regression should give a slope of 1, with all data points falling precisely upon the regression line. Since we have no prior expectation as to which variable should be on the x- versus y-axis—we are simply testing for a slope of 1—we used standardized major axis regression using the "smatr" package in R. We performed standardized major axis

regressions of *usage count(codon)*, *high occupancy windows*~*usage count(codon)*, *low occupancy windows* along with package tests that the slope of the line is 1 and that the intercept falls through 0. When we consider only those codons within the lowest quartile of frequency values, we find that the resulting regression has a slope not significantly different from 1 ($p = 0.51$) and an intercept not significantly different from 0 ($p = 0.68$), indicating that on the whole the rarest (tAI) quartile of codons are used equally between the slow and quickly translated windows. Considering all codons, however, gives a regression with both a slope different from 1 ($p = 2.9e\text{-}04$) and an intercept different from 0 ($p = 4.4e\text{-}04$), corroborating that not rarer but more common codons are used more in the high-occupancy windows. The line $x = y$ is plotted just as a visual aid. (B) An examination of the residuals from (A). Those codons that lie more than ~2 standard deviations away from the regression line are not from the rare end of the frequency spectrum but do tend to encode positively charged residues. Horizontals at $y = -1.96, +1.96$ are plotted. (C) Given that there will of course be constraints on amino acid sequence, we also desire to investigate the differences in codon usage between the two windows given the protein-coding composition of each. All of the total codon counts for each low-occupancy window (as described above) were divided by the total amino acid count encoded by that codon for the low-occupancy window. The same normalization was performed for the high-occupancy windows, and the normalized codon counts were then plotted against one another. Performing a standard major axis regression on the amino acid-adjusted codon counts shows that codons, given the protein coding sequence, are on the whole used proportionally between the quickly and slowly translated windows. When we consider only those codons within the lowest quartile of frequency values, we find that the resulting regression has a slope not significantly different from 1 ($p = 0.74$) and an intercept not significantly different from 0 ($p = 0.25$), indicating that on the whole the rarest (frequency) quartile of codons are used equally between the slow and quickly translated windows. Considering all codons, we find a slope significantly different from, but very close to, 1 ($p = 0.049$; slope 95% CI of 1.00, 1.08) and an intercept not different from 0 ($p = 0.10$). The line $x = y$ is plotted as a visual aid. (D) The finding in (C) that codons, on the whole, are not used significantly different between the slowly and quickly translated windows (given their respective amino acid compositions) is confirmed by an analysis of the residuals. The one codon that is possibly significantly overused does not have a low genomic frequency. Horizontals at $y = -1.96, +1.96$ are plotted.
(PDF)

Figure S4 Shifting the "preceding 30 codons" window 4 codons upstream to accommodate the "back" of the ribosome still shows rare codons do not slow ribosomes. Imagining ribosomes did stop at rare (tAI) codons, the A-site would still be ~10–12 nucleotides from the end of the ribosomal footprint. To make sure we are not in fact improperly normalizing footprint counts around rare clusters by a "preceding 30" sequence that contains part of the footprints, we moved the "preceding 30 codons" window upstream by four codons (i.e., 12 nt). We achieve very similar results to those presented in the main text (see Figure 2). (A) All genes with rare codon clusters. (B) Genes with rare codon clusters that have 0 or 1 positive charge coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density, with the bulk of the effect of positive charge removed. (C) Genes with rare codon clusters that have two or more positive charges in the last 30 codon positions plotted.
(PDF)

Figure S5 Pairs, triplets, etc. of rare (low tAI) codons do not tend to slow ribosomes. The consecutive rare codons in considered codons are present between the first and second arrowheads. The mean r_{pos}/r_{prec30}, or relative change in ribosomal occupancy, at each position across aligned transcripts ± s.e.m. is plotted. The horizontal at $y = 1$ represents the null expectation that positive charges do not alter ribosomal speed—that is, that ribosomes are, on average, as frequently present before the rare codon cluster as after it. (A) All genes with rare codon clusters. (B) Genes with rare codon clusters that have 0 or 1 positive charge coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density, with the bulk of the effect of positive charges removed. (C) Genes with rare codon clusters that have two or more positive charges in the last 30 plotted codon positions.
(PDF)

Figure S6 Codons that are overused in high-ribosomal occupancy windows are not "rare" according to tAI. In the main text, we examine whether "rare" codons slow ribosomes and define "rare" as the lowest quartile of tAI values within the genome. To ensure there is not a problem with this definition, we have examined the difference in trends of codon usage at large between the two windows. (A) Tallies of all the codons used among the high-occupancy and low-occupancy windows within each gene (including the preceding five codons before each window) were kept separately. We plotted the natural log of counts for each codon in the high ribosomal occupancy window versus the natural log of counts in the low occupancy window and have color coded the codons according to their tAI (see also Figure S3 for rare codons defined according to their genomic frequency). If all codons are used equally among the slowly translated and quickly translated windows, then the regression should give a slope of 1, with all data points falling precisely upon the regression line. Since we have no prior expectation as to which variable should be on the *x*- vs. *y*-axis—we are simply testing for a slope of 1—we used standardized major axis regression using the "smatr" package in R. We performed standardized major axis regressions of *usage count(codon)*, *high occupancy windows*~*usage count(codon)*, *low occupancy windows* along with package tests that the slope of the line is 1 and that the intercept falls through 0. When we consider only those codons within the lowest quartile of tAI values, we find that the resulting regression has a slope not significantly different from 1 ($p = 0.93$) and an intercept not significantly different from 0 ($p = 0.82$), indicating that on the whole the rarest (tAI) quartile of codons are used equally between the slow and quickly translated windows. Considering all codons, however, gives a regression with both a slope different from 1 ($p = 4.0e\text{-}04$) and an intercept different from 0 ($p = 5.5e\text{-}04$), corroborating that not rarer but more common codons are used more in the high-occupancy windows. The line $x = y$ is plotted just as a visual aid. (B) An examination of the residuals from (A). Those codons that lie closest to ~2 standard deviations away from the regression line tend to encode positively charged amino acids. Horizontals at $y = -1.96, +1.96$ are plotted. (C) Given that there will of course be constraints on amino acid sequence, we also desire to investigate the differences in codon usage between the two windows given the protein-coding composition of each. All of the total codon counts for each low-occupancy window (as described above) were divided by the total amino acid count encoded by that codon for the low-occupancy window. The same normalization was performed for the high-occupancy windows, and the normalized codon counts were then plotted against one another. Performing a standard major axis regression on the amino-acid-adjusted codon counts shows that codons, given the protein coding sequence, are on the whole used proportionally between the quickly and slowly translated windows. When we consider only those codons within the lowest quartile of tAI values, we find that the resulting

regression has a slope not significantly different from 1 ($p = 0.45$) and an intercept not significantly different from 0 ($p = 0.89$), indicating that on the whole the rarest (tAI) quartile of codons are used equally between the slow and quickly translated windows. Considering all codons, we find a slope significantly different from, yet very close to 1 ($p = 0.032$; slope 95% CI of 1.00, 1.10) and an intercept again not different from 0 ($p = 0.07$; intercept 95% CI of −0.034, 0.0015). The line $x = y$ is plotted as a visual aid. (D) The finding in (C) that codons, on the whole, are not used significantly differently between the slowly and quickly translated windows (given their respective amino acid compositions) is confirmed by an analysis of the residuals. The one codon that is possibly significantly overused does not have a low tAI value. Horizontals at $y = -1.96, +1.96$ are plotted.
(PDF)

Figure S7 Similarity to Kozak sequence is not the primary cause of ribosomal slowing. Given that transcript similarity to the Shine-Dalgarno sequence has been shown to slow ribosomes in bacteria due to interactions of the sequence with components of the ribosomal RNA [17], we wondered whether translation speed in yeast might not be modulated by codon usage per se but by the ability of ribosomes to bind to transcript sequence that mirrors the eukaryotic Kozak sequence. Specifically, we wanted to determine whether codons that are in high-ribosomal occupancy windows within a gene might be more likely to correspond to the Kozak sequence (as compared to codons in low-occupancy windows within the same genes) and hence bind ribosomes, slowing translation. We first determined which codons were enriched in the Kozak sequence relative to the codon frequencies seen throughout the yeast genome at large using a simple randomization. Nucleotide frequencies at each position of the Kozak sequence in yeast were taken from Cavener and Ray 1991 [57]. To determine the frequencies of all the possible "codons" among the Kozak sequence space, we randomly created 20,000 possible Kozak sequences from the delineated nucleotide frequencies at each site in the consensus sequence. We then counted all possible triplet "codons" within each sequence, regardless of reading frame (since we assume that as the ribosome traverses RNA, it may bind the Kozak sequence regardless of the surrounding reading frame). The counts of all possible RNA triplets that we observe within our simulated sequences are the observed "codons" within the Kozak sequence. In order to determine whether or not certain codons are over- or underused in the Kozak sequence, we compare them to the counts of codons observed (again in any reading frame) across 20,000 randomized sequences derived from the basal codon frequencies in the *S. cerevisiae* genome and of the same length as the Kozak sequence. We calculate z, a measure of the over- or underusage of a particular codon within the Kozak sequence (as compared to the rest of the genome) as z_{codon} = [Observed codon count (in Kozak sequence) − Expected count (from genome frequencies)]/Expected SD of codon. We can then examine which codons are overused (i.e., with a positive z-score) in slowly translated windows relative to quickly translated windows in the same genes and ask if these codons are overrepresented among the Kozak sequence(s). If so, this would suggest that RNA sequence may be slowing ribosomes not through codon–anticodon interactions but by Kozak-similar sequences binding the ribosome. (A) Tallies of all the codons used among the high-occupancy and low-occupancy windows were kept separately. We then performed a regression of count(codon) in high occupancy windows~count(codon) in low occupancy windows. The line y = x is plotted as a visual aid. (B) Standardized residuals from the analysis in (A) are plotted against the original x values in (A). No codons that are overrepresented in the Kozak sequence (i.e., have positive z-scores) have standardized residuals greater than +1.96, implying they may be overused. The high-z codon AAA comes close to the +1.96 mark, however we note that AAA encodes a positively charged amino acid, lysine, as do AAG and CGA, which also fall near the +1.96 mark and are not overused in the Kozak sequence. Horizontals are plotted at y = 1.96, +1.96. (C) Here the codon counts used in (A) were normalized by the usage of the corresponding amino acid to investigate fluctuations in synonymous codon choice given the amino acid in the protein. We then performed a regression of count(codon)/count(corresponding amino acid) in high occupancy windows~count(codon)/count(-corresponding amino acid) in low occupancy windows. The line y = x is plotted as a visual aid. (D) Standardized residuals from (C) are plotted against the original x values. We observe that those codons that are significantly overrepresented (i.e., over +1.96 standard deviations) in the high occupancy windows (given the amino acid content) are in fact underrepresented in the Kozak sequence (with a negative z-score) compared to the genome at large. Even the AAA codon, above the +1.96 standard deviation mark in (B), is not overused when factoring in amino acid choice as shown here. We consider this confirmation of our inference that the AAA codon has a high residual in (B) on account of the amino acid it encodes, and not merely because of its similarity to Kozak sequence. For these reasons, although we cannot rule out a potential contribution to slowing, we consider that transcript similarity to the Kozak sequence cannot explain the bulk of ribosomal pausing in yeast.
(PDF)

Figure S8 Ribosomal slowing after positive charge clusters in the ribosomal footprint set taken from amino acid-starved yeast [29].
(PDF)

Figure S9 Changes in relative translation rates after rare codon clusters calculated from amino-acid-starved data [29]. Three rare codon clusters are plotted with outlier axes. (A) All genes with rare codon clusters. (B) Genes with rare codon clusters that have 0 or 1 positive charge coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density with the bulk of the effect of positive charge removed. (C) Genes with rare codon clusters that have two or more positive charges in the last 30 codon positions plotted.
(PDF)

Figure S10 Positive charges show an additive (linear) trend in slowing ribosomes in the amino-acid-starved dataset [29], but rare codons do not. The degree of slowing is a function of both the magnitude of ribosomal density and the length of transcript the slowing covers. Therefore to measure any trend in the ability of either positive charges or codon clusters to slowing, the area between the curves depicting the average relative change in ribosomal density (r_{pos}/r_{prec30}) and the $y = 1$ null in Figures S8 and S9, whether positive or negative, was summed between $x = 0$ (the beginning of the cluster) and the point where the plotted values intersect with $y = 1$ again (see Figure 1). A positive value for the area under the curve indicates ribosomal slowing, while a negative value reflects faster movement. (A) Regression of *area under curve~size of cluster +0* gives a slope of 5.15 ($p = 0.0122$, $r^2 = 0.7815$). A linear model (not shown) that does not force the regression through the origin gives an insignificant intercept ($p = 0.64$). (B–D) Regression of *area under curve~size of cluster + 0*, slope $p = 0.56$, 0.93, and 0.55, respectively.
(PDF)

Figure S11 Positive charges encoded by A/G- and C-rich codons both slow ribosomes. If positive charges indeed slow

codons, we should detect slowing regardless of the codon encoding the charge. Since we are now considering specific subgroups among the positive charge clusters depending on the corresponding codon composition, sample size quickly becomes an issue. The 1-positive charge clusters give not only the best sample size, but also the fairest comparison since the composition of the "cluster" must be binary (either A/G- or C-rich) and not mixed. Our results show that positive charge slows ribosomes regardless of the nature of the codon encoding the charge. The C-rich codons (encoding Arg and His) may slow translation slightly less than the A-rich codons (Lys and Arg). This is to be expected, as histidine has a lesser tendency to be charged at physiological pH (see also Results).
(PDF)

Figure S12 Histidine-enriched clusters slow less than histidine-free clusters. As we note in the main text, histidine is less likely to be charged at physiological pH than lysine or arginine. Here we divide positive charge clusters according to whether or not they contain a minimal number of histidine residues versus no histidines at all and observe that greater slowing is observed after histidine-free clusters, in line with expectations if charge does slow ribosomes.
(PDF)

Figure S13 The only significantly overused amino acid in the high-ribosomal occupancy windows across genes (relative to the amino acid content in the paired low-occupancy windows in the same genes) is lysine, which is positively charged. In our main analysis we identified amino acids we expect to slow ribosomes (e.g., basic amino acids) and then examining the change in ribosomal occupancy upon their addition to the peptide chain. An alternative approach is to ask which amino acids are statistically overrepresented within the most slowly translated (i.e., most footprint-dense) regions within a gene. As different genes have their own expression levels, nucleotide contents, and functions, we would ideally like to control for these differences among genes when examining which amino acids are overused on the whole. For this reason, we re-employed a two-window analysis in which the highest ribosomal occupancy window and the lowest occupancy window (each of 10 codons) were identified in every gene for which we had ribosomal occupancy data. Tallies of all the amino acids used among the high-occupancy and low-occupancy windows (and including the preceding five codons before each window, as these amino acids may have just entered the tunnel when slowing occurs) were kept separately. We then performed a regression of *usage count(aa)*, *high occupancy windows*~*usage count(aa)*, *low occupancy windows*: if all amino acids are used equally among the slowly translated and quickly translated windows, then the regression should give a slope of 1, with all data points falling precisely upon the regression line. We plotted the residuals of this regression against the low window count, such that amino acids that are significantly overused in the high-occupancy window will have standardized residuals of greater than +1.96. Only a positively charged amino acid (lysine) is significantly overused in the higher ribosomal occupancy window.
(PDF)

Figure S14 The effect of positive charge is not explained by covariance with codon usage or mRNA folding. In order to determine if global patterns of codon usage or mRNA secondary structure might in fact be contributing to patterns in ribosomal slowing we observe after clusters of positive charges, we also examined the relative changes in tAI and PARS values after the clusters. Within a given transcript, the relative increase or decrease in codon optimality at each position surrounding the charged cluster was calculated by dividing the measured ribosomal density at some codon position (tAI_{pos}) (i.e., at some position before/after the charged residue is added) by the average tAI of the 30 codons preceding the first coded-for charge in the cluster within that transcript (tAI_{prec30}). The mean relative change in tAI after a cluster positive charges was then calculated by aligning all transcripts with a given cluster size by the first charge in each cluster and calculating the average ratio (tAI_{pos}/tAI_{prec30}) in each codon site surrounding the cluster. We similarly calculated the relative increase or decrease in propensity for double-stranded structure, as quantified by PARS values, at each position surrounding the charged cluster. As PARS values as originally published [38] are logged ratios, we first took the antilog of all PARS values (making all of them positive) in order to be able to calculate relative increases or decreases in the values along transcripts by dividing the antilogged PARS value at some codon position surrounding the encoded charge cluster ($PARS_{pos}$) by the average PARS of the 30 codons (all previously antilogged) preceding the first coded-for charge in the cluster within that transcript ($PARS_{prec30}$). This method is conservative, as taking the antilog will result in PARS values indicating single-strandedness being sandwiched between 0 and 1, but with PARS values indicating double-strandedness spread above 1. Hence increases in double-stranded propensity will be exaggerated. The average relative change in either tAI or PARS (mean tAI_{pos}/tAI_{prec30} or $PARS_{pos}/PARS_{prec30}$) at a given position after a cluster was then calculated by aligning all identified regions of a given cluster size according to the first charge present in each cluster and calculating the average ratio in positions increasingly distant from the first positive charge of the aligned clusters. Positive charges in a cluster may be coded for anywhere between the two downturned triangles. An average r_{pos}/r_{prec30} above 1 indicates a relative local increase in ribosomal density in that position across transcripts (as in Figure 1). (A) An average tAI_{pos}/tAI_{prec30} below 1 indicates the codons in that position across transcripts tend to decrease in optimality on average relative to the average tAI of the preceding 30 codons across transcripts, while a ratio above 1 signifies an increase in optimality. We find that differential codon use in the vicinity of positive charges cannot explain the charge slowing effect. We observe no correlation between relative changes in ribosomal density and tAI after the first charge in the cluster ($0 \geq x \leq 30$ in this figure, panel A; Spearman P, left to right: 0.93, 0.73, 0.22, 0.17, and 0.65). For a more relaxed test, we then compared, for each plot in Figure 5, the relative changes in codon optimality (tAI_{pos}/tAI_{prec30}) seen after the start of each cluster at $x = 0$ until the point where relative change in ribosomal density (r_{pos}/r_{prec30}) drops back to previous levels ($y = 1$) to the tAI_{pos}/tAI_{prec30} values seen in all other surrounding plotted sites (i.e., those sites lacking charge-induced pausing). If anything, relatively more optimal ($tAI_{pos}/tAI_{prec30} > 1$) codons are coded for during periods of elevated ribosomal occupancy for clusters comprising six or more encoded cations, while no difference in optimality is detected in codon usage during elevated ribosomal occupancy compared to surrounding codon usage for other-sized charge clusters (Mann-Whitney U test p values, left to right in this figure, panel A: 0.96, 0.20, 0.07, 0.07, and 0.003). Hence we conclude that changes in codon bias are not responsible for the slowing patterns associated with positively charged residues (Figure 5), as expected if rare codons do not slow ribosomes (Figure 3A,B). (B) An average relative change in (here antilogged, see Methods) PARS values (i.e., $PARS_{pos}/PARS_{prec30}$) plotted above 1 indicates a greater likelihood of double-stranded structure in that position on average relative to preceding sequence, while a ratio less than 1 indicates a decrease in propensity for double-strandedness relative to the

preceding 30 codons. We find that the slowing effect of positive charge cannot be explained by mRNA folding in the vicinity of positive charges. There is no correlation between the relative change in PARS values ($PARS_{pos}/PARS_{prec30}$) after the first charge in the cluster (this Figure, panel B, $0 \geq x \leq 30$) and relative changes in ribosomal density (Spearman P, left to right: 0.44, 0.68, 0.97, 0.99, and 0.15), which we may have expected to observe if RNA structure has a local effect on ribosomal slowing. Likewise, under such a local-slowing hypothesis, we should expect to see a significant difference in the average PARS ratios seen among the sequence between $x = 0$ and the point at which elevated ribosomal density curve (r_{pos}/r_{prec30}) drops back to $y = 1$ versus PARS ratios in surrounding plotted sites. Such a difference, however, is seen only in the two-charge plot (this figure, panel B; Mann-Whitney U test p values, left to right: 0.17, 0.0006, 0.24, 0.08, and 0.60). If we instead assume that downstream structure has a pausing effect observable more upstream, a more appropriate test is to compare the PARS ratios from $-30 \geq x < 0$ to those from $0 \geq x \leq 30$. In this case, we observe no significant difference in relative propensity for double-strandedness before or after positive charges apart from in the case of a single positive charge alone [this figure, panel B; Mann-Whitney U test, left to right: 0.004, 0.07, 0.12, 0.08 (with the mean $PARS_{pos}/PARS_{prec30}$ decreasing on average after the start of the cluster), and 0.60]. We note that this version of the test is exceedingly conservative as PARS values had to be antilogged before informative ratios could be calculated. This means that previously negative values (indicating single-strandedness) will now be sandwiched in between 0 and 1, while formerly positive values (indicating double-strandedness) now span a range of values above 1. Hence normalizing the PARS score at a given position by the average PARS value of the preceding 30 codons will exaggerate not only the importance of structured versus free-form RNA, but will also exaggerate small differences in the magnitude of PARS values already denoting double-strandedness. (C) An alternative calculation showing that RNA structure does not account for the pausing observed near positive charges. Note this figure does not show the change in PARS values relative to the preceding sequence (as in B), but the average magnitude of the PARS value in that position across aligned transcripts. An average of PARS values plotted above 0 indicates a greater likelihood of double-stranded structure in that position on average, while a mean value of less than 1 indicates a propensity for single-strandedness. We find no correlation between the average PARS values after the first charge in the cluster ($0 \geq x \leq 30$) and relative changes in ribosomal density (this figure, panel C; Spearman P, left to right: 0.77, 0.95, 0.87, 0.34, and 0.09), as we might have observed if RNA structure has a local effect on ribosomal slowing. Likewise, if structure causes local slowing, we should see a significant difference in the average PARS values between $x = 0$ and the point at which elevated ribosomal density curve (r_{pos}/r_{prec30}) drops back to $y = 1$ versus PARS values in surrounding plotted sites. We do not, however, detect such a difference (this figure, panel C; Mann-Whitney U test p values, left to right: 0.66, 0.17, 0.30, 0.27, and 0.90). Examining whether downstream structure has a pausing effect observable further upstream, we then compare the PARS ratios from $-30 \geq x < 0$ to those from $0 \geq x \leq 30$. In this case, we observe no significant difference in relative propensity for double-strandedness before or after positive charges (this figure, panel C; Mann-Whitney U test, left to right: 0.98, 0.98, 0.97, 0.27, and 0.90).
(PDF)

Figure S15 Genes with either high or low footprint coverage both produce consistent slowing patterns after positive charge clusters. To ensure that noise in the location of footprints among genes with fewer overall footprints is not an issue for analysis, we redrew our r_{pos}/r_{prec30} plots surrounding positive charge clusters using both the bottom half and top half of all genes according to their footprint saturation. Note that in this analysis we do not normalize the footprint counts per codon per gene by mRNA levels. This is because we are not interested in footprint coverage per transcript (as we might be if considering rates or mechanistic issues), but in the statistical power that the total footprint coverage per gene gives us, regardless of the number of transcripts that the footprints were captured from. Areas under the curve were measured as in the main text (see Figure 1). In each case we find similar results to those presented in the main analysis (Figure 5), namely that positive charges additively slow ribosomes. (A) Bottom half of genes: Regression of *area under curve*~*cluster size*, slope = 4.8, $r^2 = 0.79$, $p = 0.027$. (B) Top half of genes: Regression of *area under curve*~*cluster size*, slope = 0.96, $r^2 = 0.74$, $p = 0.039$.
(PDF)

Figure S16 Figure 5 redone on the nonredundant footprint set. We wanted to confirm that the exclusion of footprints that map to two or more potential locations in the genome was not systematically biasing our estimates of ribosomal density. For this reason we replotted the average relative change in ribosomal density within a gene upon translation of encoded positive charge clusters using our nonredundant footprint set (see the end of the Methods section), in effect only considering those locations in the genome to which footprints uniquely map. Considering solely these regions in the transcriptome to which footprints can only ever be mapped unambiguously still shows positive charges additively slow translation.
(PDF)

Figure S17 Figure 2 redone on the nonredundant footprint set. We wanted to confirm that the exclusion of footprints that map to two or more potential locations in the genome was not systematically biasing our estimates of ribosomal density. For this reason we replotted the average relative change in ribosomal density within a gene upon translation of rare codon clusters using our nonredundant footprint set (see the end of the Methods section), in effect only considering those locations in the genome to which footprints uniquely map. Considering solely these regions in the transcriptome to which footprints can only ever be mapped unambiguously still shows rare codons do not slow translation. (A) All genes with rare codon clusters. (B) Genes with rare codon clusters that have 0 or 1 positive charge coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density with the bulk of the effect of positive charge removed. (C) Genes with rare codon clusters that have two or more positive charges in the last 30 codon positions plotted.
(PDF)

Note S1 Codon usage and translation rates: how can codon usage not predict ribosome occupancy but be commonly assumed to be associated with faster translation?
(PDF)

Note S2 Only positive charge is capable of explaining the region of strongest translational pausing within transcripts.
(PDF)

Note S3 Trend of slowing increasing with charge is not random.
(PDF)

Table S1 Table 1 of the main text redone using rare codons that are defined to occur with genomic infrequency shows rare codons do not slow ribosomes. In the main text, we investigate whether nonoptimal codons—that is, those with low tAI scores—might slow codons and find that they do not. To ensure that our finding

that these "rare" codons do not slow ribosomes does not simply hinge on our definition of "rare," we have repeated the analysis using an alternative definition. Here, we define "rare" codons according to their actual frequency in the genome as measured from our set of filtered genes. This rare set, of equal size to the rare tAI set, comprises the following codons: CGG, CGC, CGA, TGC, CCG, CTC, GGG, GCG, CGT, CCC, CAC, TGT, ACG, TCG, and AGG. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. Rare (infrequent) codons and codon pairs tend to be found more in the less dense (faster translated) window. Similarly, the presence of rare pairs and rare codons decreases in the slowly translated windows as the difference in degree of ribosomal slowing grows.
(PDF)

Table S2 Genes with identified rare codon clusters are not disproportionately sampled from lowly expressed genes. Could it be that large changes in ribosomal occupancy are not observed after rare clusters (Figure 2A and Figure 3A) because the clusters we identify are more likely to come from lowly expressed genes—that is, genes that do not have high translation levels and for which it may be less likely that ribosomal footprints will be sampled? We used the average footprint count of a gene (total number of footprints within the coding sequence divided by gene length) as a proxy for protein expression levels. If anything, there are more genes with nonoptimal codon clusters from genes that have more footprint reads (χ^2, $p<2.2e\text{-}16$) so we do not consider this an issue.
(PDF)

Table S3 Sequence similarity to the yeast Kozak sequence cannot explain the greatest slowing within transcripts. Given that transcript similarity to the Shine-Dalgarno sequence has been shown to slow ribosomes in bacteria due to interactions of the sequence with components of the ribosomal RNA [17], we wondered whether translation speed in yeast might not be modulated by codon usage per se but by the ability of ribosomes to bind to transcript sequence that mirrors the eukaryotic Kozak sequence. Specifically, we wanted to determine whether codons that are in high-ribosomal occupancy windows within a gene might be more likely to correspond to the Kozak sequence (as compared to codons in low-occupancy windows within the same genes) and hence bind ribosomes, slowing translation. We first determined which codons were enriched in the Kozak sequence relative to the codon frequencies seen throughout the yeast genome at large using a simple randomization. Nucleotide frequencies at each position of the Kozak sequence in yeast were taken from Cavener and Ray 1991 [57]. To determine the frequencies of all the possible "codons" among the Kozak sequence space, we randomly created 20,000 possible Kozak sequences from the delineated nucleotide frequencies at each site in the consensus sequence. We then counted all possible triplet "codons" within each sequence, regardless of reading frame (since we assume that as the ribosome traverses RNA, it may bind the Kozak sequence regardless of the surrounding reading frame). The counts of all possible RNA triplets that we observe within our simulated sequences are the observed "codons" within the Kozak sequence. In order to determine whether or not certain codons are over- or underused in the Kozak sequence, we compare them to the counts of codons observed (again in any reading frame) across 20,000 randomized sequences derived from the basal codon frequencies in the *S. cerevisiae* genome and of the same length as the Kozak sequence. We calculate $\mathcal{Z}$, a measure of the over- or underusage of a particular codon within the Kozak sequence (as compared to the rest of the genome) as $\mathcal{Z}_{\text{codon}}$ = [Observed codon count (in Kozak sequence) − Expected count (from genome frequencies)]/Expected SD of codon. We can then perform a test similar to the one in Methods, "The Relative Contributions of Charge, Folding, and Codon Usage to Extremes of Slowing Within Transcripts," but where we consider possible slowing codons to be those with a positive $\mathcal{Z}$ (GAT GAC AAC TGC CAA GGC GTA GTC TAT ACA TGG ATA CAT AAA TGT AAT ATG). A score of 1 indicates there are more codons with positive $\mathcal{Z}$ within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. (A) Similarity to Kozak sequence cannot explain slowing in several quantiles (binomial tests), nor can it explain increased slowing (χ^2 tests). (B) Even when the number of positive charges is the same between the two windows, we do not detect a significant contribution of similarity to Kozak sequence to slowing. (C) Controlling for amino acid usage in two different ways, we detect no contribution of similarity to Kozak sequence to slowing; in fact, as the degree of slowing increases, the ability of Kozak similarity to explain slowing decreases (χ^2 tests). Method 1 (in bold): a gene is scored "1" if the slow window contains more codons with positive $\mathcal{Z}$, "−1" if it contains fewer. Method 2 (in italics): the magnitude of all the positive $\mathcal{Z}$ values is averaged in each window, and the gene is scored "1" if the slower window has a higher average $\mathcal{Z}$, "−1" if its average $\mathcal{Z}$ is lower.
(PDF)

Table S4 Table 1 tAI score tests controlled for amino acid content. Could differences in amino acid usage between the two windows be biasing our result that neither codon usage nor rare pairs slow ribosomes (Table 1)? It could be that certain amino acids only have relatively high or low tAIs, and a preponderance of such amino acids in one window over the other could cause an apparent preference for (non-)optimal codons, which is in fact a preference for a certain amino acid. For this reason we tested whether differences in amino acid usage between the high and low ribosomal occupancy windows within a transcript systematically alter the tAI scores (and hence the resulting interpretation of the contribution of codon usage to ribosomal density) in our window comparison analysis. To do this, we identified the same high and low ribosomal occupancy windows within a transcript as above. This time, however, we considered only amino acids that are coded for at least once within each window. Within each intra-transcript window, we identified all codons that code for amino acid *x* and quantified the contribution of tAI to ribosomal occupancy using two approaches: (Method 1) The average tAI of all the codons coding for amino acid (aa) *x* was calculated for each window, and that amino acid was assigned an aa-tAIscore of 1, 0, or −1, depending on whether the tAI in the higher ribosomal occupancy window was lower (and hence capable of explaining the increased ribosomal density), the same, or higher than that in the other window, respectively. All of the aa-tAI scores for a given gene were counted independently—in other words, for a given gene, it was possible to calculate more than one aa-tAI score, and all these aa-tAI scores contributed to the final matrix. (Method 2) The average tAI of all the codons coding for amino acid *x* in each window was calculated, similarly to Method 1, but a tAI score is not yet assigned. Instead, the average tAI is first determined for each amino acid present in both windows, and then average tAIs (each the average for a particular amino acid) are themselves averaged to come up with a single aa-tAI for each window. Then, a single tAI score is assigned to that gene by comparing the average aa-tAIs in each window similarly to above. Bold, Method 1; italic,

Method 2. Original Δr quantiles means the same quantile boundaries used in the main analysis were used, whereas recalculated Δr quantiles are drawn from only those genes considered in this amino-acid-adjusted analysis. The p value for χ^2 tests with fewer than five observations in any square was calculated by resampling the observations without replacement and noting how many times (r) the χ^2 value of the resampled set was greater than or equal to the observed. The p was then calculated as $(r+1)/(n+1)$, where n is the number of iterations performed (1,000). (A) Upon controlling for differential amino acid content in the two windows as detailed above, the result that tAI cannot explain patterns of slowing is still robust. Additionally we no longer detect a decrease in the ability of tAI to explain pausing in the upper quantiles as observed in Table 1A. (B) and (C) show the effect of tAI (adjusted for amino acid use) in only those pairs of intra-transcript windows that have the same number of positive charges between them.
(PDF)

Table S5 Table 1 done again on the amino-acid-starved footprint set [29]. Only positive charge is systematically capable of explaining ribosomal slowing, including the severest slowing. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. A low codon optimality, if anything, tends to pair more with the less dense (faster translated) window. Similarly, not only do rare pairs and rare 6-mers tend to be found more often in the faster-translated window, but their presence decreases as the difference in degree of ribosomal slowing grows. Additionally, a greater likelihood of transcript secondary structure at or just before the identified window is associated not with the more occluded windows, but with the less dense (faster translated) ones, and the presence of secondary structure in fact decreases as the difference in ribosomal slowing between the windows increases. Positive charge, however, is consistently associated with the higher density (more slowly translated) window.
(PDF)

Table S6 Positive charge best explains the slowest translated regions within transcripts compared to other physiochemical properties of amino acids. While we find that positive charges slow ribosomes, we wanted to control for the effects of other physiochemical properties of amino acids, specifically hydropathy (Phe, Val, Leu, Ile, Met), polarity (Asn, Gln, Ser, Thr, Cys, Tyr), and negative charge (Asp, Glu). These groups of amino acids, however, do not lend themselves to the r_{pos}/r_{prec30} analysis we carry out in the main text (see Figures 1–5) in the same way that positive charge does. The r_{pos}/r_{prec30} plotting analysis is suited to positive charges because they cluster in a way that gives us reasonable sample sizes given our constraints—that is, the number of positive charges we require in the cluster and the additional requirement that there be no surrounding positive charges outside of the cluster. In the case of the other amino acid groups, there are either too many constituent members of the group and which are used too frequently (e.g., hydropathy) to define isolated "clusters" for investigation, or the amino acids are used too rarely as clusters away from positive charges, and are of insufficient cluster sizes to establish any slowing trends (e.g., negative charges). We therefore compared the effects of these other physiochemical properties of amino acids by comparing the amino acids encoded by the highest ribosomally occupied versus lowest occupied windows within genes. The analysis was carried out similarly to the way Table 1 was created in the main text, only this time counting different amino acids depending on the physiochemical property being investigated. We find that, on the whole, only positive charge can robustly explain the slowing patterns we observe. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. (A) Hydrophobic residues (Phe, Val, Leu, Ile, Met) cannot explain increased slowing as the difference in translation speed between the two windows increases (χ^2 $p = 0.98$). Additionally the proportion of genes that pass the hydrophobicity test compared to failing it is only significant in the fourth quantile (q4) (binomial $p = 0.023$). (B) Polar residues (Asn, Gln, Ser, Thr, Cys, Tyr) cannot explain increased slowing as the difference in translation speed between the two windows increases (χ^2 $p = 0.21$). Additionally the proportion of genes that pass the polarity test compared to failing it is only significant in the fourth quantile (q4) (binomial $p = 3.7e\text{-}08$). (C) Negative charges (Asp, Glu) cannot explain increased slowing as the difference in translation speed between the two windows increases (χ^2 $p = 0.14$). Additionally the number of genes that pass or fail the negative charge score test in the third quantile (q3) is not significantly different (binomial $p = 0.83$). (D) Positive charge score, from Table 1, is shown for purposes of comparison.
(PDF)

Table S7 Positive charge explains slowing better than amino acid hydrophobicity. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. (A–C) In those genes that fail the positive charge test (charge score = 0 or −1), we find that hydrophobicity cannot explain the increased slowing in these windows either (this table, χ^2 tests). For this reason we consider that while amino acids with hydrophobic side chains may be used more often in the vicinity of positive charge (this table, binomial tests), perhaps for certain structural motifs or because of the types of genes under consideration, they cannot be responsible for the major slowing effect. (D–F) Positive charge can explain the slowing in genes where hydrophobicity cannot.
(PDF)

Table S8 Positive charge explains slowing better than amino acid polarity. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. (A–C) In those genes that fail the positive charge test (charge score = 0 or −1), we find that polarity cannot explain the increased slowing in these windows either (this table, χ^2 tests). For this reason we consider that while amino acids with polar side chains may be used more often in the vicinity of positive charge (this table, binomial tests), perhaps for certain structural motifs or because of the types of genes under consideration, they cannot be responsible for the major slowing effect. (D–F) Positive charge can explain the slowing in some genes where polarity cannot.
(PDF)

Table S9 Positive charge explains slowing better than negative charge. Quantiles of the difference in average ribosomal density

between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. (A–C) In those genes that fail the positive charge test (charge score = 0 or −1), we find that negatively charged amino acids cannot explain the increased slowing in these windows either (this table, χ^2 tests). For this reason we consider that while amino acids with negatively charged side chains may be used more often in the vicinity of positive charge (this table, binomial tests), perhaps for certain structural motifs or because of the types of genes under consideration, they cannot be responsible for the major slowing effect. (D–F) Positive charge can explain the slowing in genes where negative charge cannot.
(PDF)

Table S10 The relationship of charge score to tAI score. Quantiles of the difference in average ribosomal occlusion between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. (A) The ability of charge to explain slowing (charge score of 1) cannot be explained by concomitant use of suboptimal codons. A charge score of 1 more commonly pairs with a tAI score, which cannot explain slowing (tAI score of −1), and increasingly so as the difference in ribosomal speeds between the two windows grows. (B) These tAI scores are drawn from transcripts for which both intra-transcript windows have the same number of charges (charge score = 0) and hence such comparisons should be controlled for the effect of positive charge on ribosomal speed. Different tAI scores are equally distributed among quantiles, indicating the inability of tAI to predict either ribosomal slowing or the degree of ribosomal slowing even in the absence of an effect of charge on ribosomal speed. (C) tAI does not systematically account for slowing in windows for which increased charge pairs with the faster window.
(PDF)

Table S11 The relationship of rare pair score to charge score. Quantiles of the difference in average ribosomal occlusion between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts. (A) The ability of charge to explain slowing (charge score of 1) cannot be explained by concomitant use of rare pairs. A charge score of 1, if anything, tends to pair with a rare pair score that cannot explain slowing (rare pair score of −1). (B) These rare pair scores are drawn from transcripts for which both intra-transcript windows have the same number of charges (charge score = 0) and hence such comparisons should be controlled for the effect of positive charge on ribosomal speed. Different rare pair scores are equally distributed among quantiles, indicating the inability of rare pairs to predict ribosomal slowing. Additionally, as the difference in the degree of ribosomal slowing increases (i.e., moving from q1 to q4), the number of rare pairs found in the higher occupancy window decreases (χ^2 test), demonstrating rare pairs cannot predict the magnitude of slowing even in the absence of an effect of charge on ribosomal speed. (C) Rare pairs do not systematically account for slowing in windows for which increased charge pairs with the faster window.
(PDF)

Table S12 The relationship of PARS score (double strandedness) to charge score. Quantiles of the difference in average ribosomal occlusion between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; −1, less present; 0, present in both windows in equal amounts.
(PDF)

Table S13 Table 1 done again, allowing the lower occupancy window to have a ribosomal occupancy of 0.
(PDF)

Table S14 Table 1 done again on the non-redundant footprint location set.
(PDF)

Author Contributions

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: LDH CAC. Performed the experiments: LDH CAC. Analyzed the data: LDH CAC. Contributed reagents/materials/analysis tools: LDH CAC. Wrote the paper: LDH CAC.

References

1. Randall LL, Josefsson LG, Hardy SJ (1980) Novel intermediates in the synthesis of maltose-binding protein in Escherichia coli. Eur J Biochem 107: 375–379.
2. Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. J Mol Biol 396: 1310–1318.
3. Doma MK, Parker R (2006) Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. Nature 440: 561–564.
4. Yanofsky C (1981) Attenuation in the control of expression of bacterial operons. Nature 289: 751–758.
5. Chartrand P, Meng XH, Huttelmaier S, Donato D, Singer RH (2002) Asymmetric sorting of ash1p in yeast results from inhibition of translation by localization elements in the mRNA. Mol Cell 10: 1319–1330.
6. Mariappan M, Li X, Stefanovic S, Sharma A, Mateja A, et al. (2010) A ribosome-associating factor chaperones tail-anchored membrane proteins. Nature 466: 1120–1124.
7. Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151: 389–409.
8. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, et al. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315: 525–528.
9. Anderson WF (1969) The effect of tRNA concentration on the rate of protein synthesis. Proc Natl Acad Sci U S A 62: 566–573.
10. Gouy M, Gautier C (1982) Codon usage in bacteria—correlation with gene expressivity. Nucleic Acids Res 10: 7055–7074.
11. Thanaraj TA, Argos P (1996) Ribosome-mediated translational pause and protein domain organization. Protein Sci 5: 1594–1612.
12. Cortazzo P, Cervenansky C, Marin M, Reiss C, Ehrlich R, et al. (2002) Silent mutations affect in vivo protein folding in Escherichia coli. Biochem Biophys Res Commun 293: 537–541.
13. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9: r43–r74.
14. Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257: 3026–3031.
15. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147: 789–802.
16. Qian W, Yang J-R, Pearson NM, Maclean C, Zhang J (2012) Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet 8: e1002603. doi:10.1371/journal.pgen.1002603
17. Li GW, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538–541.
18. Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136: 927–935.

19. Stoletzki N, Eyre-Walker A (2007) Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol 24: 374–381.
20. Precup J, Parker J (1987) Missense misreading of asparagine codons as a function of codon identity and context. J Biol Chem 262: 11351–11355.
21. Warnecke T, Hurst LD (2010) GroEL dependency affects codon usage-support for a critical role of misfolding in gene evolution. Molecular Systems Biology 6: 340
22. Wen JD, Lancaster L, Hodges C, Zeri AC, Yoshimura SH, et al. (2008) Following translation by single ribosomes one codon at a time. Nature 452: 598–603.
23. Somogyi P, Jenner AJ, Brierley I, Inglis SC (1993) Ribosomal pausing during translation of an RNA pseudoknot. Mol Cell Biol 13: 6931–6940.
24. Kozak M (1986) Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. Proc Natl Acad Sci U S A 83: 2850–2854.
25. Lu J, Kobertz WR, Deutsch C (2007) Mapping the electrostatic potential within the ribosomal exit tunnel. J Mol Biol 371: 1378–1391.
26. Lu J, Deutsch C (2008) Electrostatics in the ribosomal tunnel modulate chain elongation rates. J Mol Biol 384: 73–86.
27. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, et al. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141: 344–354.
28. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, et al. (2011) Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12: R110.
29. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218–223.
30. Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A 107: 3645–3650.
31. Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129: 897–907.
32. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research 32: 5036–5044.
33. Kane JF (1995) Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. Curr Opin Biotechnol 6: 494–500.
34. Varenne S, Baty D, Verheij H, Shire D, Lazdunski C (1989) The maximum rate of gene expression is dependent on the downstream context of unfavourable codons. Biochimie 71: 1221–1229.
35. Frenkel-Morgenstern M, Danon T, Christian T, Igarashi T, Cohen L, et al. (2012) Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Mol Syst Biol 8: 572.
36. Brackley CA, Romano MC, Thiel M (2011) The dynamics of supply and demand in mRNA translation. PLoS Comput Biol 7: e1002203. doi:10.1371/journal.pcbi.1002203
37. Elf J, Nilsson D, Tenson T, Ehrenberg M (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. Science 300: 1718–1722.
38. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, et al. (2010) Genome-wide measurement of RNA secondary structure in yeast. Nature 467: 103–107.
39. Ito-Harashima S, Kuroha K, Tatematsu T, Inada T (2007) Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. Genes Dev 21: 519–524.
40. Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158: 573–597.
41. Curran JF, Yarus M (1989) Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J Mol Biol 209: 65–77.
42. Nakatogawa H, Ito K (2002) The ribosomal exit tunnel functions as a discriminating gate. Cell 108: 629–636.
43. Bhushan S, Meyer H, Starosta AL, Becker T, Mielke T, et al. (2010) Structural basis for translational stalling by human cytomegalovirus and fungal arginine attenuator peptide. Mol Cell 40: 138–146.
44. Fang P, Spevak CC, Wu C, Sachs MS (2004) A nascent polypeptide domain that can regulate translation elongation. Proc Natl Acad Sci U S A 101: 4059–4064.
45. Brown CE, Sachs AB (1998) Poly(A) tail length control in Saccharomyces cerevisiae occurs by message-specific deadenylation. Mol Cell Biol 18: 6548–6559.
46. Meaux S, Van Hoof A (2006) Yeast transcripts cleaved by an internal ribozyme provide new insight into the role of the cap and poly(A) tail in translation and mRNA decay. RNA 12: 1323–1337.
47. Inada T, Aiba H (2005) Translation of aberrant mRNAs lacking a termination codon or with a shortened 3′-UTR is repressed after initiation in yeast. EMBO J 24: 1584–1595.
48. Akimitsu N, Tanaka J, Pelletier J (2007) Translation of nonSTOP mRNA is repressed post-initiation in mammalian cells. EMBO J 26: 2327–2338.
49. Dimitrova LN, Kuroha K, Tatematsu T, Inada T (2009) Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. J Biol Chem 284: 10343–10352.
50. Gillet R, Felden B (2001) Emerging views on tmRNA-mediated protein tagging and ribosome rescue. Mol Microbiol 42: 879–885.
51. Bengtson MH, Joazeiro CA (2010) Role of a ribosome-associated E3 ubiquitin ligase in protein quality control. Nature 467: 470–473.
52. Percudani R, Pavesi A, Ottonello S (1997) Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol 268: 322–330.
53. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32: D493–D496.
54. R Development Core Team (2005) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
55. Lu J, Deutsch C (2005) Folding zones inside the ribosomal exit tunnel. Nat Struct Mol Biol 12: 1123–1129.
56. Yonath A, Leonard KR, Wittmann HG (1987) A tunnel in the large ribosomal subunit revealed by three-dimensional image reconstruction. Science 236: 813–816.
57. Cavener DR, Ray SC (1991) Eukaryotic start and stop translation sites. Nucleic Acids Res 19: 3185–3192.

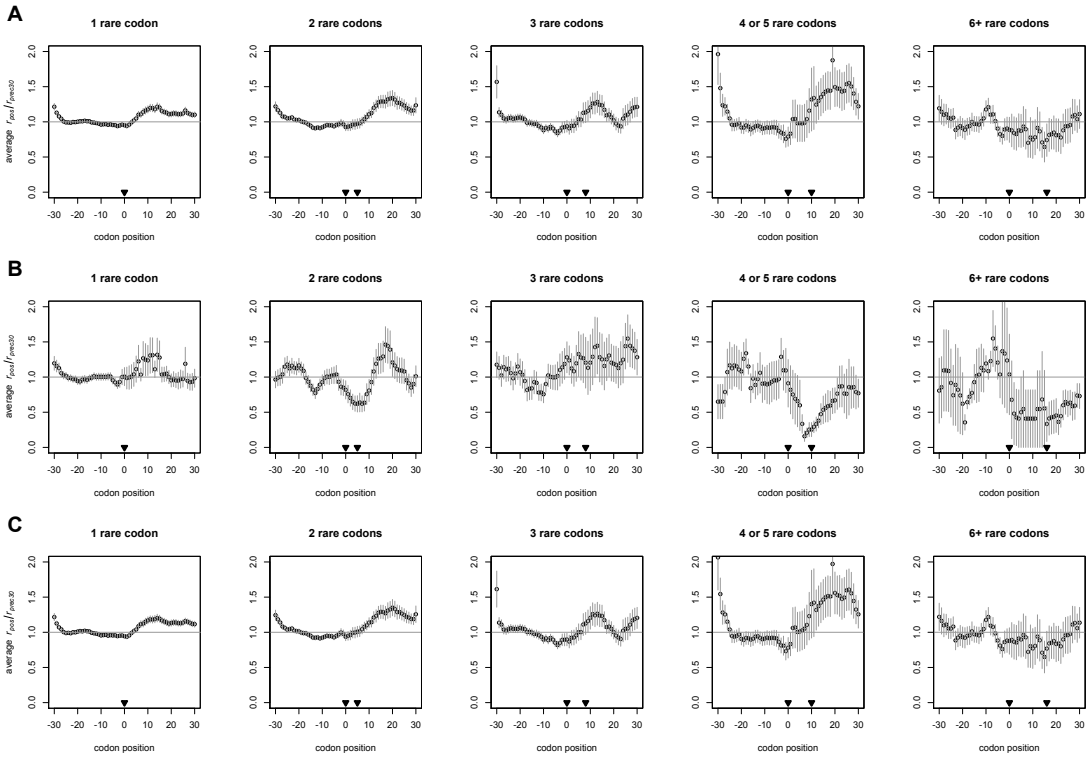

Figure S1. Figure 2 redone using rare codons defined according to genomic frequency shows rare codons do not slow ribosomes. In the main text we investigate whether non-optimal codons, i.e. those with low tAI scores, might slow codons and find that they do not. To ensure that our finding that these 'rare' codons do not slow ribosomes does not simply hinge on our definition of 'rare', we have repeated the analysis using an alternative definition. Here, we define 'rare' codons according to their actual frequency in the genome as measured from our set of filtered genes. This rare set, of equal size to the rare tAI set, comprises the following codons: CGG, CGC, CGA, TGC, CCG, CTC, GGG, GCG, CGT, CCC, CAC, TGT, ACG, TCG, AGG. We find that rare codons, where rare means genomically rare, do not slow ribosomes when in clusters (single rare codons; two rare codons anywhere within a 5-codon stretch; 3 rare codons within 8 codons; 4 or 5 within 10, and 6 or more within 16). Note slowing should be observed over, not after, the rare codon(s). **A)** All genes with rare codon clusters. Regression of *area under curve ~ number of rare codons in cluster*, slope = -0.79, P = 0.080. Regressions were performed as detailed in the main text (see Figure 1 for a description of the calculation of area under the curve). We note even if P were significant, the slope would be negative, whereas if rare codons did slow ribosomes we should expect to see a positive slope. **B)** Genes with rare codon clusters which have 0 or 1 positive charges coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density with the bulk of the effect of positive charge removed. **C)** Genes with rare codon clusters which have 2 or more positive charges in the last 30 codon positions plotted.

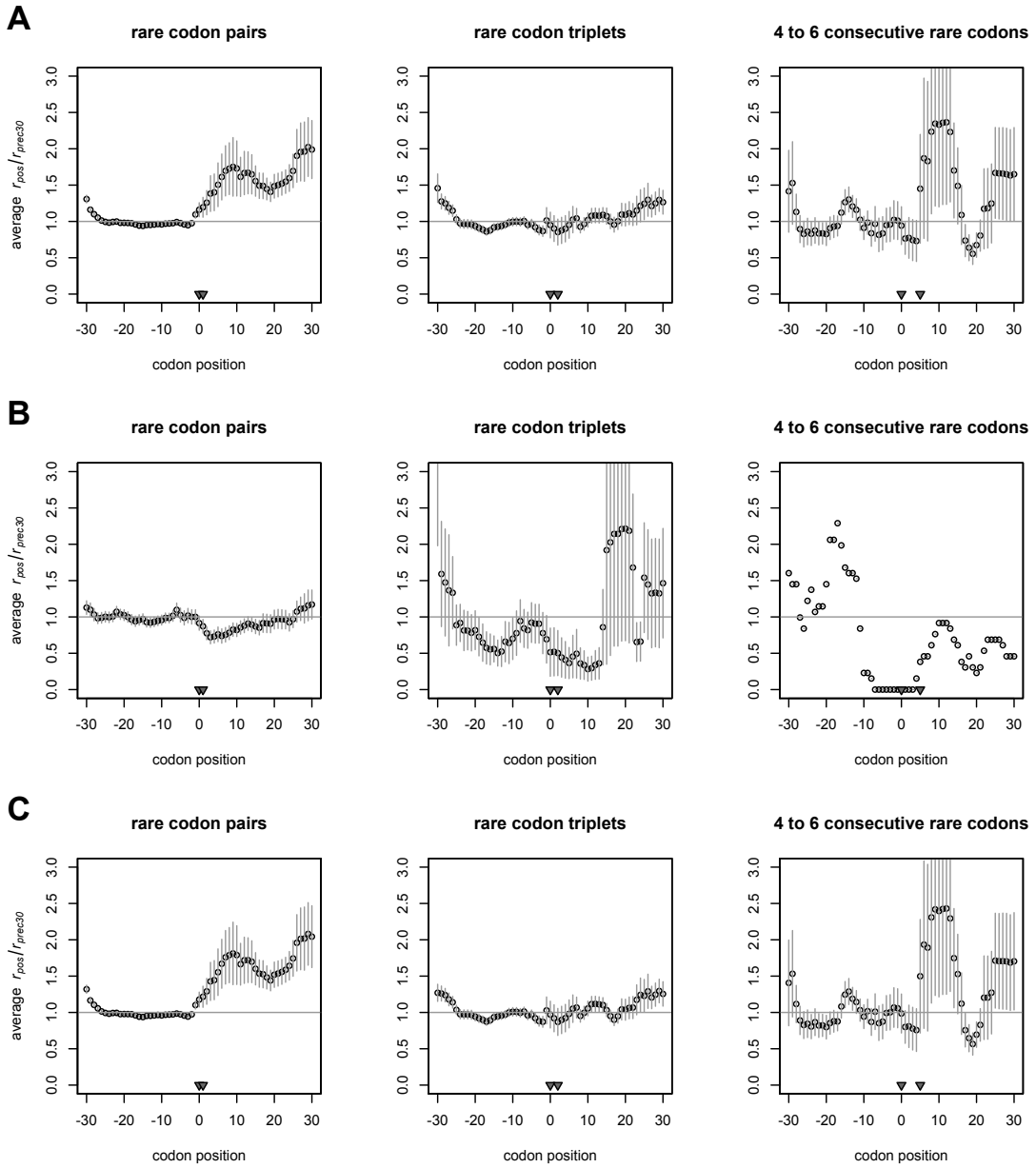

Figure S2. Consecutive rare codons, where rare is with reference to genomic frequency, do not slow ribosomes. In the main text we investigate whether non-optimal codons, i.e. those with low tAI scores, might slow codons and find that they do not. To ensure that our finding that these 'rare' codons do not slow ribosomes does not simply hinge on our definition of 'rare', we have repeated the analysis using an alternative definition. Here, we define 'rare' codons according to their actual frequency in the genome as measured from our set of filtered genes. This rare set, of equal size to the rare tAI set, comprises the following codons: CGG, CGC, CGA, TGC, CCG, CTC, GGG, GCG, CGT, CCC, CAC, TGT, ACG, TCG, AGG. The consecutive rare codons in considered codons are present between the first and second arrowheads. See Figure 1 for a description of the calculation of the area under the curve. **A)** All genes with rare codon clusters. Regression of *area under curve ~ number of rare codons in cluster*, slope = -9.8, P = 0.32. **B)** Genes with rare codon clusters which have 0 or 1 positive charges coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density with the bulk of the effect of positive charge removed. **C)** Genes with rare codon clusters which have 2 or more positive charges in the last 30 plotted codon positions.

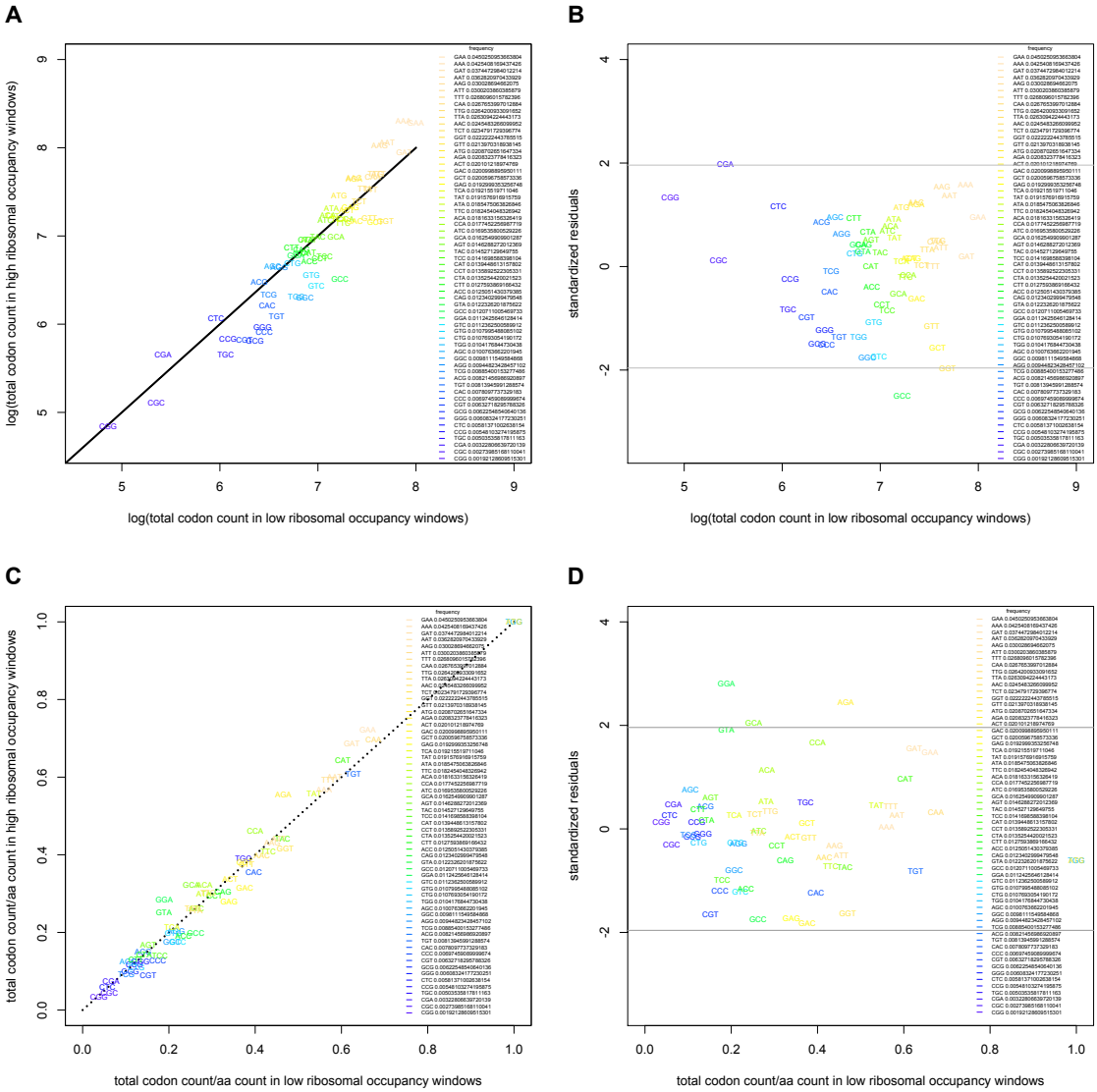

Figure S3. Codons which are overused in high-ribosomal occupancy windows are not 'rare' according to genomic frequency. In some supplemental analyses we examine whether 'rare' codons slow ribosomes, and define 'rare' as the quartile of those most infrequent codons in the genome. To ensure there is not a problem with this definition, we have examined the difference in trends of codon usage at large between the two windows. **A.** Tallies of all the codons used among the high-occupancy and low-occupancy windows within each gene (including the preceding 5 codons before each window) were kept separately. We plotted the counts for each codon in the high ribosomal occupancy window versus the counts in the low occupancy window, and have color-coded the codons according to their frequency (see also Figure S6 for rare codons defined according to their tAI). If all codons are used equally among the slowly-translated and quickly-translated windows then the regression should give a slope of 1, with all datapoints falling precisely upon the regression line. Since we have no prior expectation as to which variable should be on the *x*- vs. *y*-axis—we are simply testing for a slope of 1—we used standardized major axis regression using the 'smatr' package in R. We performed standardized major axis regressions of *usage count(codon), high occupancy windows ~ usage count(codon), low occupancy windows* along with package tests that the slope of the line is 1 and that the intercept falls through 0. When we consider only those codons within the lowest quartile of frequency values, we find that the resulting regression has a slope not significantly different from one (P = 0.51) and an intercept not significantly different from 0 (P = 0.68), indicating that on the whole the rarest (tAI) quartile of codons are used equally between the slow and quickly-translated windows.

Considering all codons, however, gives a regression with both a slope different from 1 (P = 2.9e-04) and an intercept different from 0 (P = 4.4e-04), corroborating that not rarer but more common codons are used more in the high-occupancy windows. The line $x = y$ is plotted just as a visual aid. **B.** An examination of the residuals from part A. Those codons which lie more than ~2 standard deviations away from the regression line are not from the rare end of the frequency spectrum but do tend to encode positively charged residues. Horizontals at y = -1.96, +1.96 are plotted. **C.** Given that there will of course be constraints on amino acid sequence, we also desire to investigate the differences in codon usage between the two windows given the protein-coding composition of each. All of the total codon counts for each the low-occupancy window (as described above) were divided by the total amino acid count encoded by that codon for the low-occupancy window. The same normalization was performed for the high-occupancy windows, and the normalized codon counts were then plotted against one another. Performing a standard major axis regression on the amino acid-adjusted codon counts shows that codons, given the protein coding sequence, are on the whole used proportionally between the quickly and slowly-translated windows. When we consider only those codons within the lowest quartile of frequency values, we find that the resulting regression has a slope not significantly different from one (P = 0.74) and an intercept not significantly different from 0 (P = 0.25), indicating that on the whole the rarest (frequency) quartile of codons are used equally between the slow and quickly-translated windows. Considering all codons, we find a slope significantly different from, but very close to, 1 (P =0.049; slope 95% CI of 1.00, 1.08) and an intercept not different from 0 (P = 0.10). The line $x = y$ is plotted as a visual aid. **D.** The finding in part C that codons, on the whole, are not used significantly differently between the slowly and quickly translated windows (given their respective amino acid compositions) is confirmed by an analysis of the residuals. The one codon which is possibly significantly over-used is does not have a low genomic frequency. Horizontals at y = -1.96, +1.96 are plotted.

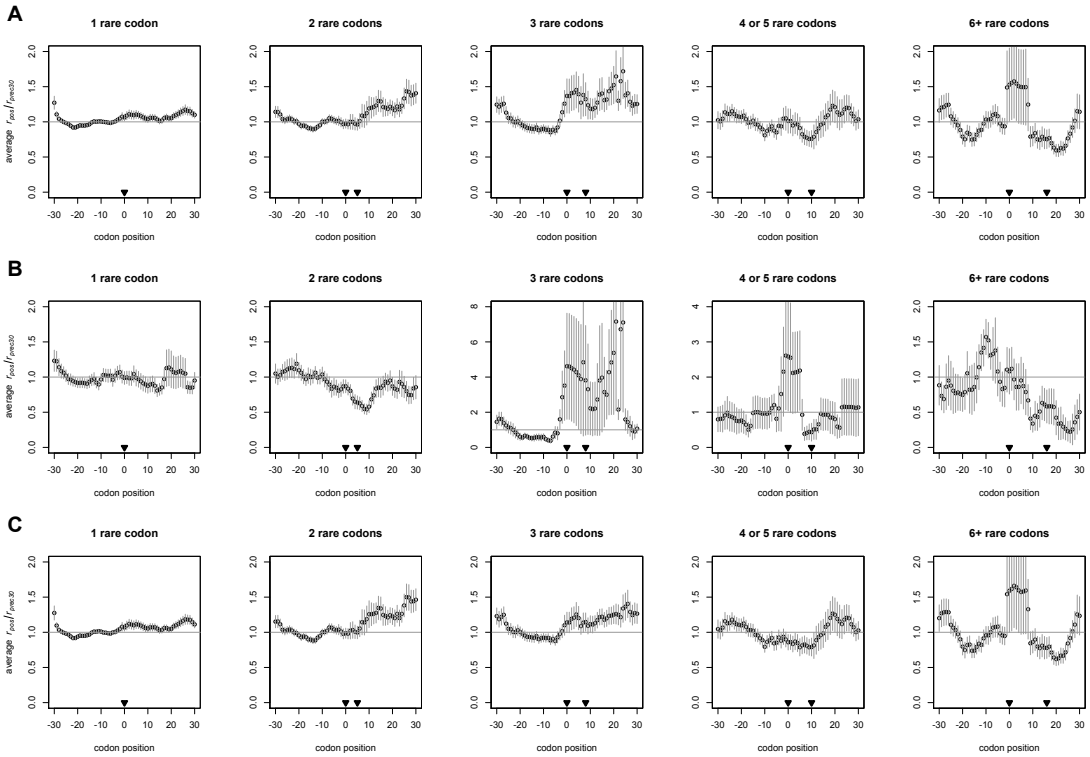

Figure S4. Shifting the 'preceding 30 codons' window 4 codons upstream to accommodate the 'back' of the ribosome still shows rare codons do not slow ribosomes. Imagining ribosomes did stop at rare (tAI) codons, the A-site would still be ~10-12 nucleotides from the end of the ribosomal footprint. To make sure we are not in fact improperly normalizing footprint counts around rare clusters by a 'preceding 30' sequence which contains part of the footprints, we moved the 'preceding 30 codons' window upstream by 4 codons (i.e. 12 nt). We achieve very similar results to those presented in the main text (see Figure 2). **A)** All genes with rare codon clusters. **B)** Genes with rare codon clusters which have 0 or 1 positive charges coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density with the bulk of the effect of positive charge removed. **C)** Genes with rare codon clusters which have 2 or more positive charges in the last 30 codon positions plotted.

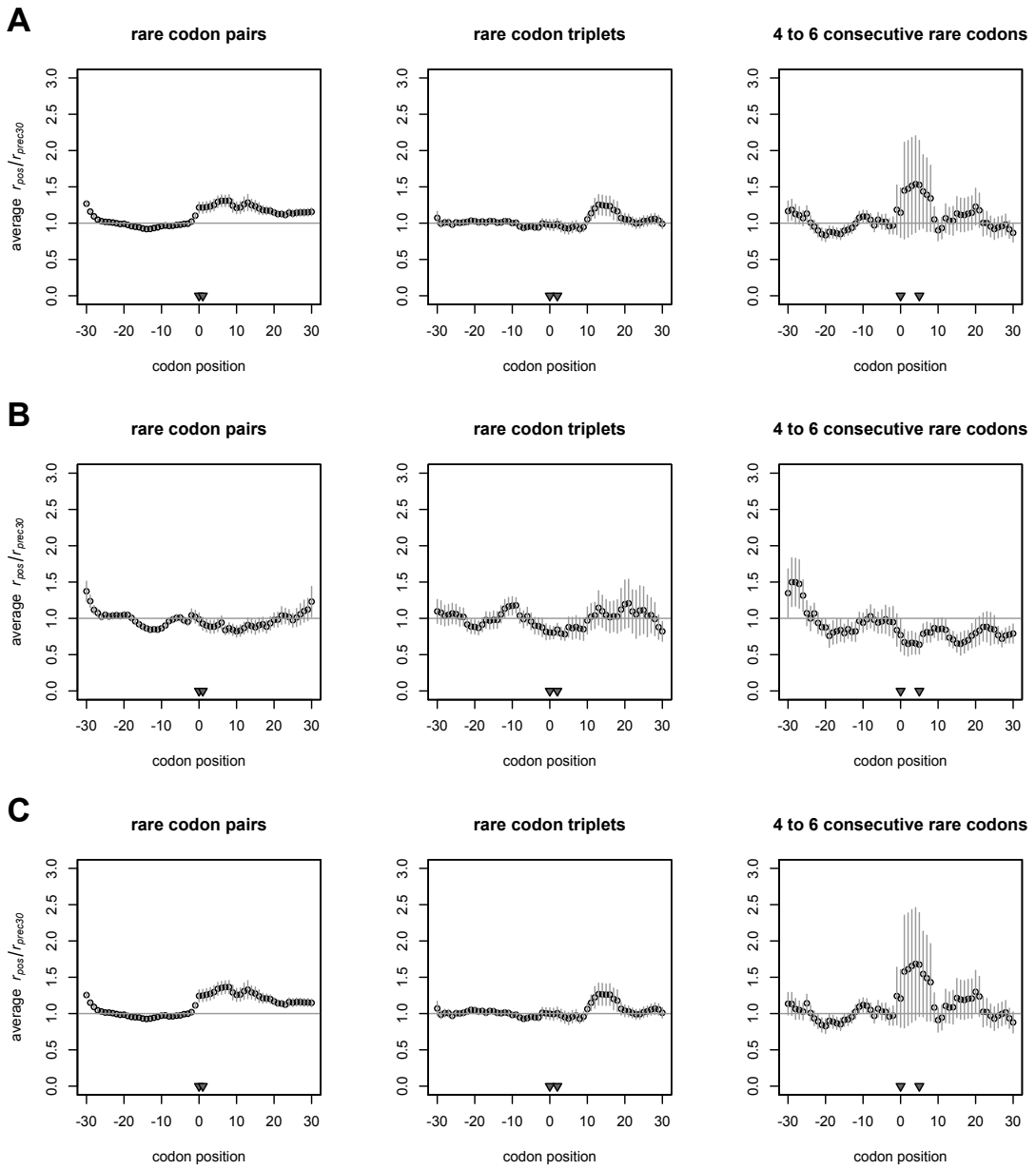

Figure S5. Pairs, triplets etc. of rare (low tAI) codons do not tend to slow ribosomes. The consecutive rare codons in considered codons are present between the first and second arrowheads. The mean r_{pos}/r_{prec30}, or relative change in ribosomal occupancy, at each position across aligned transcripts ± s.e.m. is plotted. The horizontal at $y = 1$ represents the null expectation that positive charges do not alter ribosomal speed, i.e. that ribosomes are, on average, as frequently present before the rare codon cluster as after it. **A)** All genes with rare codon clusters. **B)** Genes with rare codon clusters which have 0 or 1 positive charges coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density with the bulk of the effect of positive charge removed. **C)** Genes with rare codon clusters which have 2 or more positive charges in the last 30 plotted codon positions.

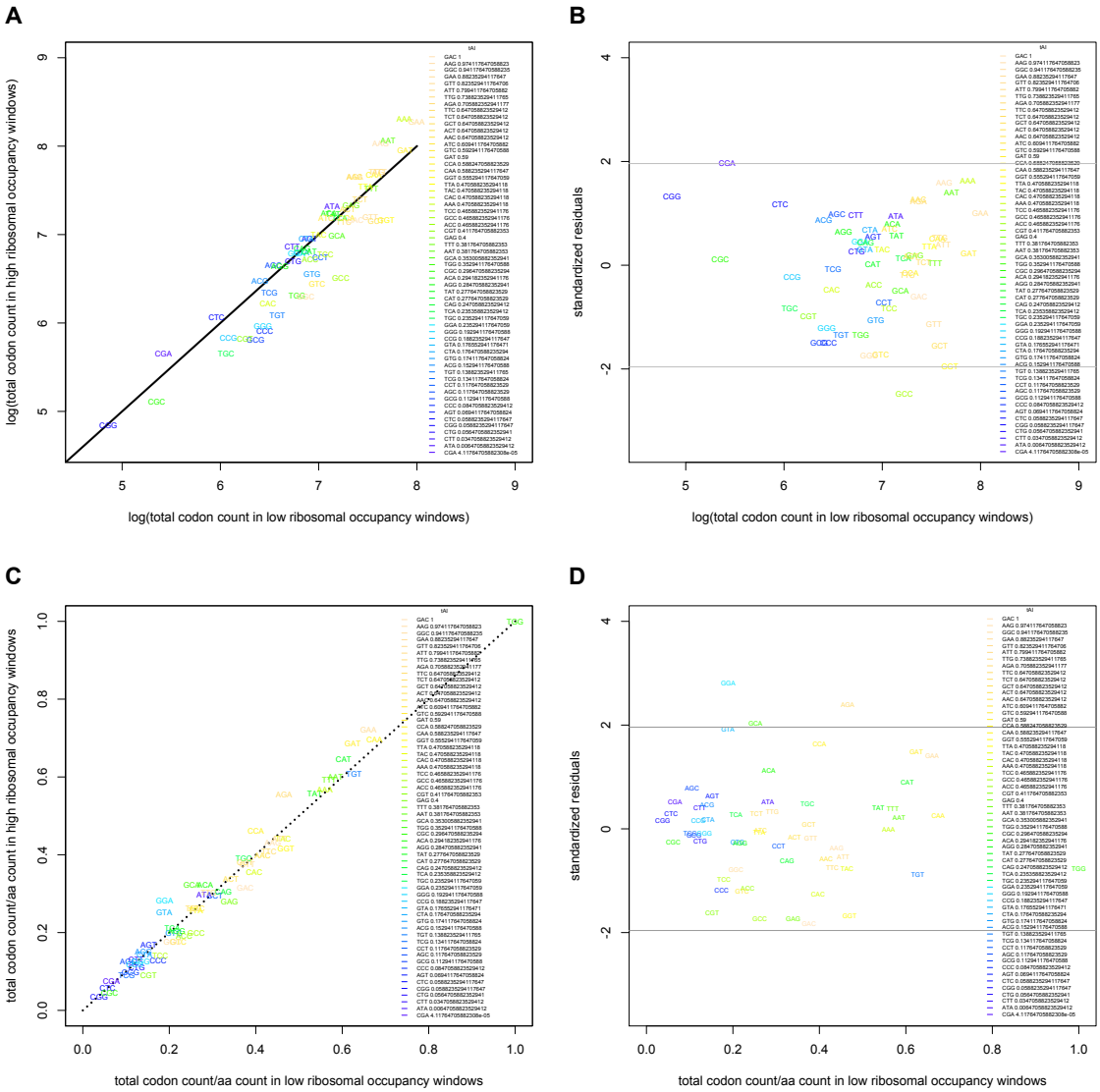

Figure S6. Codons which are overused in high-ribosomal occupancy windows are not 'rare' according to tAI. In the main text we examine whether 'rare' codons slow ribosomes, and define 'rare' as the lowest quartile of tAI values within the genome. To ensure there is not a problem with this definition, we have examined the difference in trends of codon usage at large between the two windows. **A.** Tallies of all the codons used among the high-occupancy and low-occupancy windows within each gene (including the preceding 5 codons before each window) were kept separately. We plotted the natural log of counts for each codon in the high ribosomal occupancy window versus the natural log of counts in the low occupancy window, and have color coded the codons according to their tAI (see also Figure S3 for rare codons defined according to their genomic frequency). If all codons are used equally among the slowly-translated and quickly-translated windows then the regression should give a slope of 1, with all datapoints falling precisely upon the regression line. Since we have no prior expectation as to which variable should be on the *x*- vs. *y*-axis—we are simply testing for a slope of 1—we used standardized major axis regression using the 'smatr' package in R. We performed standardized major axis regressions of *usage count(codon), high occupancy windows ~ usage count(codon), low occupancy windows* along with package tests that the slope of the line is 1 and that the intercept falls through 0. When we consider only those codons within the lowest quartile of tAI values, we find that the resulting regression has a slope not significantly different from one (P = 0.93) and an intercept not significantly different from 0 (P = 0.82), indicating that on the whole the rarest (tAI) quartile of codons are used equally between the slow and quickly-translated windows. Considering all codons, however, gives a

regression with both a slope different from 1 (P = 4.0e-04) and an intercept different from 0 (P = 5.5e-04), corroborating that not rarer but more common codons are used more in the high-occupancy windows. The line $x = y$ is plotted just as a visual aid. **B.** An examination of the residuals from part A. Those codons which lie closest to ~2 standard deviations away from the regression line tend to encode positively charged amino acids. Horizontals at y = -1.96, +1.96 are plotted. **C.** Given that there will of course be constraints on amino acid sequence, we also desire to investigate the differences in codon usage between the two windows given the protein-coding composition of each. All of the total codon counts for each the low-occupancy window (as described above) were divided by the total amino acid count encoded by that codon for the low-occupancy window. The same normalization was performed for the high-occupancy windows, and the normalized codon counts were then plotted against one another. Performing a standard major axis regression on the amino acid-adjusted codon counts shows that codons, given the protein coding sequence, are on the whole used proportionally between the quickly and slowly-translated windows. When we consider only those codons within the lowest quartile of tAI values, we find that the resulting regression has a slope not significantly different from one (P = 0.45) and an intercept not significantly different from 0 (P = 0.89), indicating that on the whole the rarest (tAI) quartile of codons are used equally between the slow and quickly-translated windows. Considering all codons, we find a slope significantly different from, yet very close to 1 (P =0.032; slope 95% CI of 1.00, 1.10) and an intercept again not different from 0 (P = 0.07; intercept 95% CI of -0.034, 0.0015). The line $x = y$ is plotted as a visual aid. **D.** The finding in part C that codons, on the whole, are not used significantly differently between the slowly and quickly translated windows (given their respective amino acid compositions) is confirmed by an analysis of the residuals. The one codon which is possibly significantly over-used is does not have a low tAI value. Horizontals at y = -1.96, +1.96 are plotted.

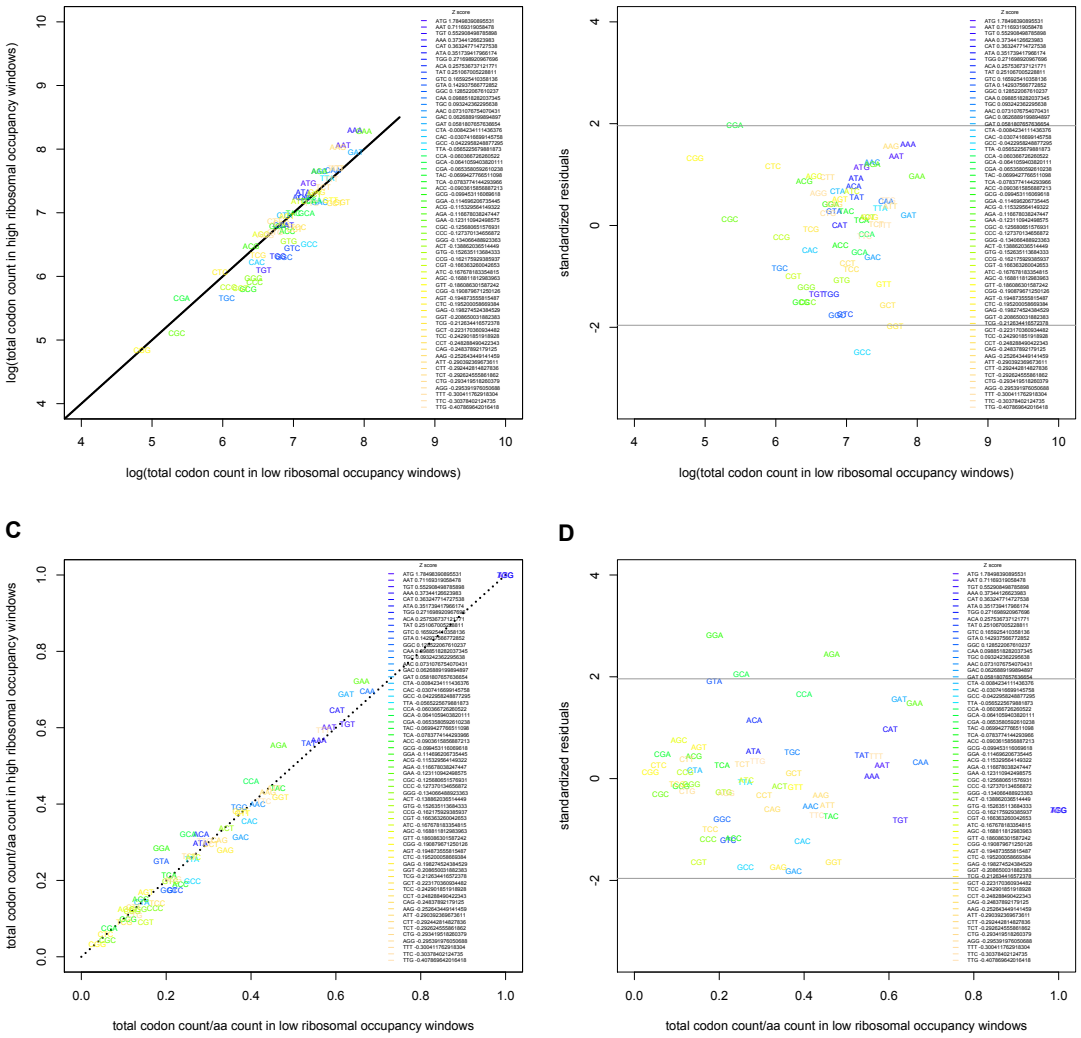

Figure S7. Similarity to Kozak sequence is not the primary cause of ribosomal slowing. Given that transcript similarity to the Shine-Dalgarno sequence has been shown to slow ribosomes in bacteria due to interactions of the sequence with components of the ribosomal RNA [17], we wondered whether translation speed in yeast might not be modulated by codon usage per se but by the ability of ribosomes to bind to transcript sequence which mirrors the eukaryotic Kozak sequence. Specifically, we wanted to determine whether codons which are in high-ribosomal occupancy windows within a gene might be more likely to correspond to the Kozak sequence (as compared to codons in low-occupancy windows within the same genes) and hence bind ribosomes, slowing translation. We first determined which codons were enriched in the Kozak sequence relative to the codon frequencies seen throughout the yeast genome at large using a simple randomization. Nucleotide frequencies at each position of the Kozak sequence in yeast were taken from Cavener and Ray 1991 [57]. To determine the frequencies of all the possible 'codons' among the Kozak sequence space, we randomly created 20000 possible Kozak sequences from the delineated nucleotide frequencies at each site in the consensus sequence. We then counted all possible triplet 'codons' within each sequence, regardless of reading frame (since we assume that as the ribosome traverses RNA, it may bind the Kozak sequence regardless of the surrounding reading frame). The counts of all possible RNA triplets that we observe within our simulated sequences are the observed 'codons' within the Kozak sequence. In order to determine whether or not certain codons are over- or under-used in the Kozak sequence, we compare them to the counts of codons observed (again in any reading frame) across 20000 randomized sequences derived from the basal codon frequencies in the S. cerevisiae genome and of the same

length as the Kozak sequence. We calculate *Z*, a measure of the over- or under-usage of a particular codon within the Kozak sequence (as compared to the rest of the genome) as Z_{codon} = [Observed codon count (in Kozak sequence) – Expected count (from genome frequencies)] / Expected SD of codon. We can then examine which codons are over-used (i.e. with a positive *Z*-score) in slowly-translated windows relative to quickly-translated windows in the same genes and ask if these codons are overrepresented among the Kozak sequence(s). If so, this would suggest that RNA sequence may be slowing ribosomes not through codon:anticodon interactions but by Kozak-similar sequences binding the ribosome. **A.** Tallies of all the codons used among the high-occupancy and low-occupancy windows were kept separately. We then performed a regression of count(codon) in high occupancy windows ~ count(codon) in low occupancy windows. The line y = x is plotted as a visual aid. **B.** Standardized residuals from the analysis in part A are plotted against the original x values in A. No codons which are over-represented in the Kozak sequence (i.e. have positive *Z*-scores) have standardized residuals greater than +1.96, implying they may be overused. The high-*Z* codon AAA comes close to the + 1.96 mark, however we note that AAA encodes a positively charged amino acid, lysine, as do AAG and CGA which also fall near the +1.96 mark and are not overused in the Kozak sequence. Horizontals are plotted at y = 1.96, +1.96. **C.** Here the codon counts used in part A were normalized by the usage of the corresponding amino acid to investigate fluctuations in synonymous codon choice given the amino acid in the protein. We then performed a regression of count(codon) / count(corresponding amino acid) in high occupancy windows ~ count(codon) / count(corresponding amino acid) in low occupancy windows. The line y = x is plotted as a visual aid. **D.** Standardized residuals from C are plotted against the original x values. We observe that those codons which are significantly over-represented (i.e. over +1.96 standard deviations) in the high occupancy windows (given the amino acid content) are in fact under-represented in the Kozak sequence (with a negative *Z*-score) compared to the genome at large. Even the AAA codon, above the +1.96 standard deviation mark in part B, is not over-used when factoring in amino acid choice as shown here. We consider this confirmation of our inference that the AAA codon has a high residual in part B on account of the amino acid it encodes, and not merely because of its similarity to Kozak sequence. For these reasons, although we cannot rule out a potential contribution to slowing, we consider that transcript similarity to the Kozak sequence cannot explain the bulk of ribosomal pausing in yeast.

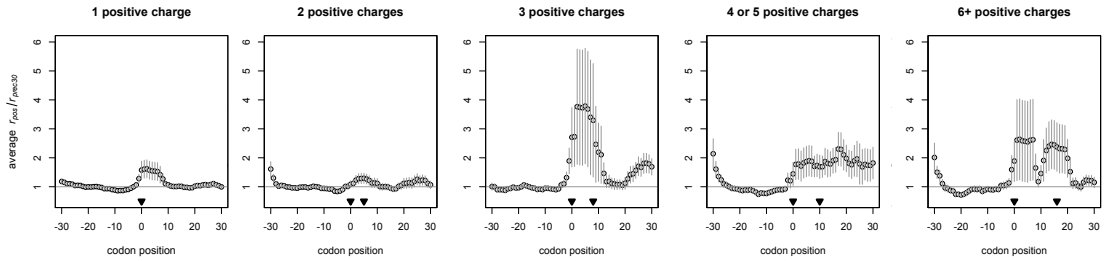

Figure S8. Ribosomal slowing after positive charge clusters in the ribosomal footprint set taken from amino acid-starved yeast [29]**.**

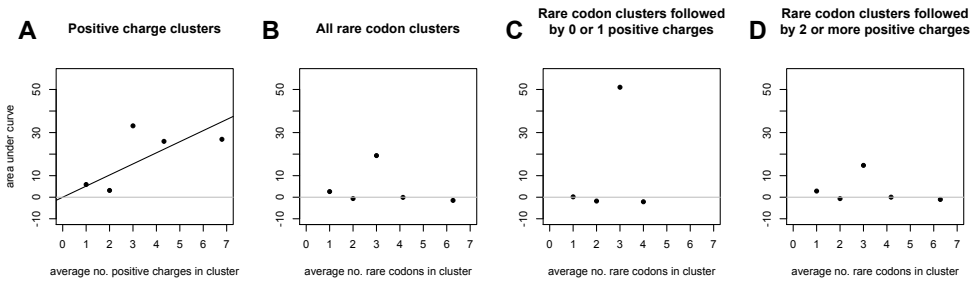

Figure S10. Positive charges show an additive (linear) trend in slowing ribosomes in the amino acid-starved dataset [29]**, but rare codons do not.** The degree of slowing is a function of both the magnitude of ribosomal density and the length of transcript the slowing covers. Therefore to measure any trend in the ability of either positive charges or codon clusters to slowing, the area between the curves depicting the average relative change in ribosomal density (r_{pos}/r_{prec30}) and the y=1 null in Figures S8 and S9, whether positive or negative, was summed between x=0 (the beginning of the cluster) and the point where the plotted values intersect with y=1 again (see Figure 1). A positive value for the area under the curve indicates ribosomal slowing, while a negative value reflects faster movement. **A)** Regression of *area under curve ~ size of cluster +0* gives a slope of 5.15 (P = 0.0122, r^2=0. 0.7815). A linear model (not shown) that does not force the regression through the origin gives an insignificant intercept (P=0.64). **B, C, D)** Regression of *area under curve ~ size of cluster + 0,* slope P = 0.56, 0.93, 0.55, respectively.

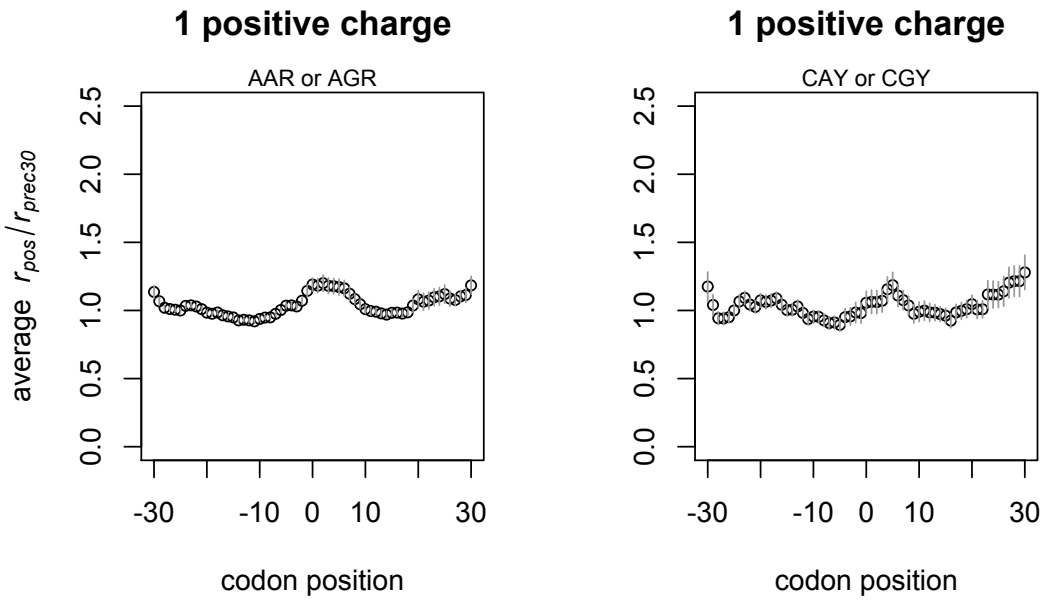

Figure S11. Positive charges encoded by A/G- and C-rich codons both slow ribosomes. If positive charges indeed slow codons, we should detect slowing regardless of the codon encoding the charge. Since we are now considering specific subgroups among the positive charge clusters depending on the corresponding codon composition, sample size quickly becomes an issue. The 1-positive charge clusters give not only the best sample size but also the fairest comparison since the composition of the 'cluster' must be binary (either A/G- or C-rich) and not mixed. Our results show that positive charge slows ribosomes regardless of the nature of the codon encoding the charge. The C-rich codons (encoding Arg and His) may slow translation slightly less than the A-rich codons (Lys and Arg). This is to be expected as histidine has a lesser tendency to be charged at physiological pH (see also Results).

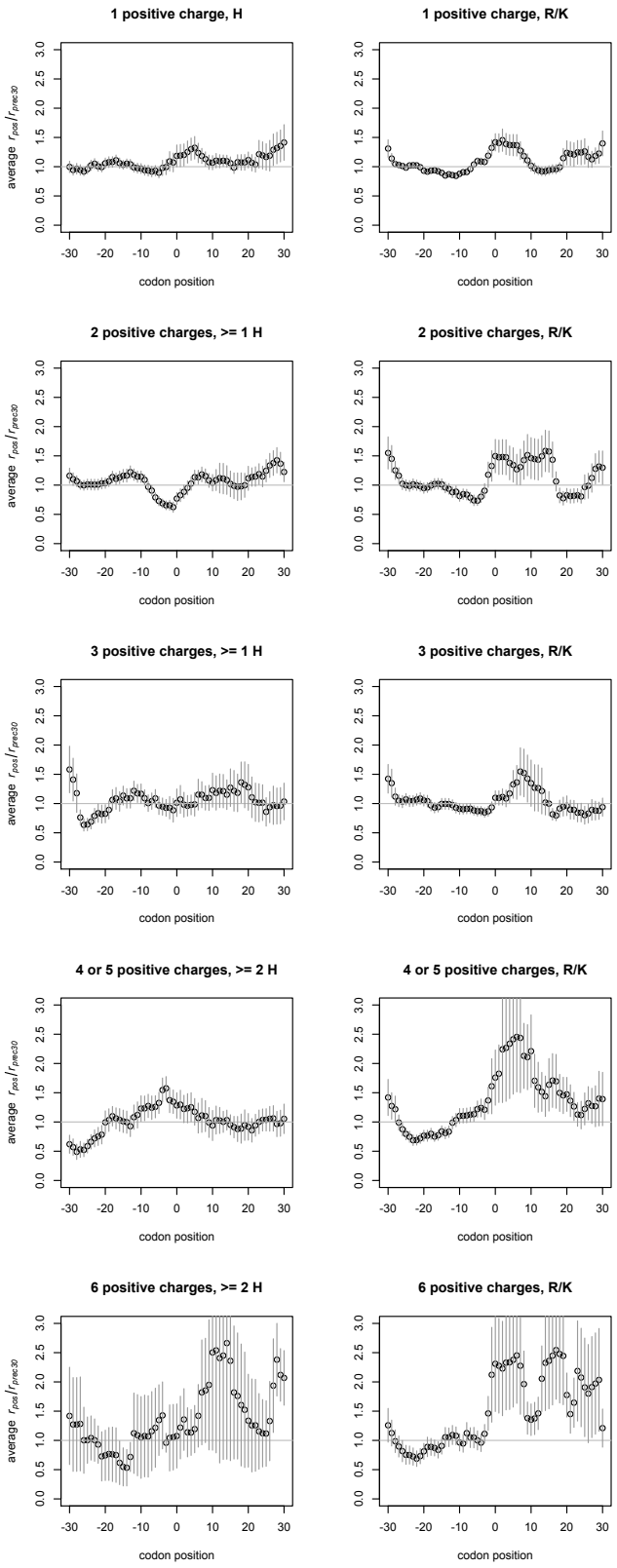

Figure S12. Histidine-enriched clusters slow less than histidine-free clusters. As we note in the main text, histidine is less likely to be charged at physiological pH than lysine or arginine. Here we divide positive charge clusters according to whether or not they contain a minimal number of histidine residues versus no histidines at all and observe that greater slowing is observed after histidine-free clusters, in line with expectations if charge does slow ribosomes.

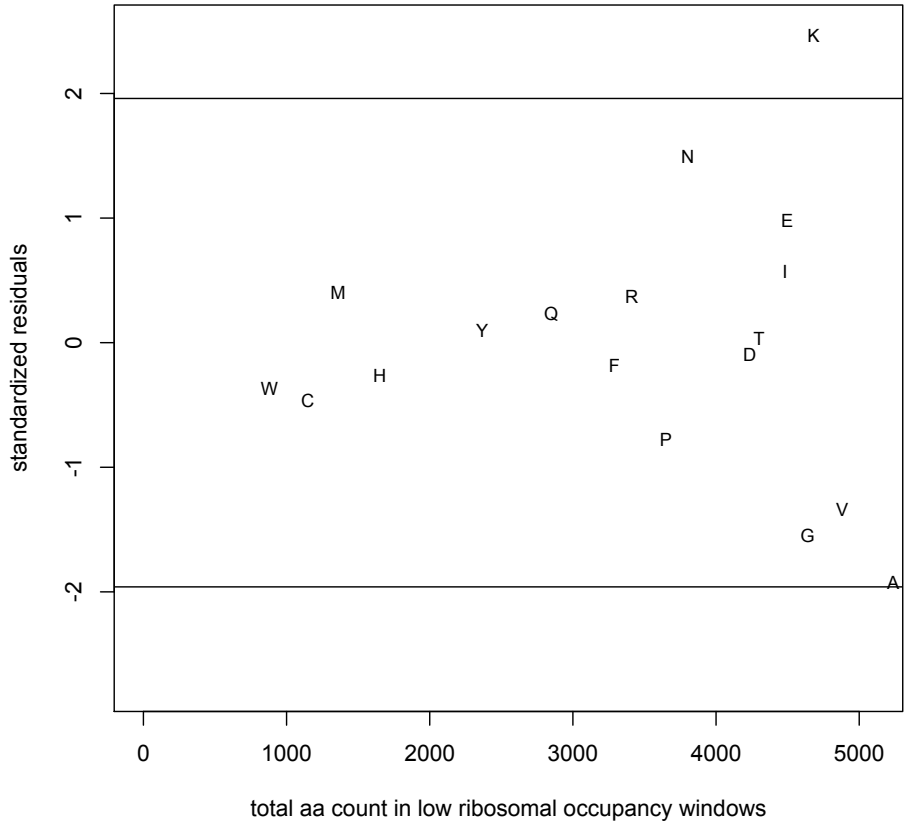

Figure S13. The only significantly overused amino acid in the high-ribosomal occupancy windows across genes (relative to the amino acid content in the paired low-occupancy windows in the same genes) is lysine, which is positively charged. In our main analysis we identified amino acids we expect to slow ribosomes (e.g. basic amino acids) and then examining the change in ribosomal occupancy upon their addition to the peptide chain. An alternative approach is to ask which amino acids are statistically overrepresented within the most slowly translated (i.e. most footprint-dense) regions within a gene. As different genes have their own expression levels, nucleotide contents, and functions, we would ideally like to control for these differences among genes when examining which amino acids are overused on the whole. For this reason we re-employed a two-window analysis in which the highest ribosomal occupancy window and the lowest-occupancy window (each of 10 codons) were identified in every gene for which we had ribosomal occupancy data. Tallies of all the amino acids used among the high-occupancy and low-occupancy windows (and including the preceding 5 codons before each window, as these amino acids may have just entered the tunnel when slowing occurs) were kept separately. We then performed a regression of *usage count(aa), high occupancy windows ~ usage count(aa), low occupancy windows*: if all amino acids are used equally among the slowly-translated and quickly-translated windows then the regression should give a slope of 1, with all datapoints falling precisely upon the regression line. We plotted the residuals of this regression against the low window count, such that amino acids which are significantly overused in the high-occupancy window will have standardized residuals of greater than +1.96. Only a positively-charged amino acid (lysine) is significantly overused in the higher ribosomal occupancy window.

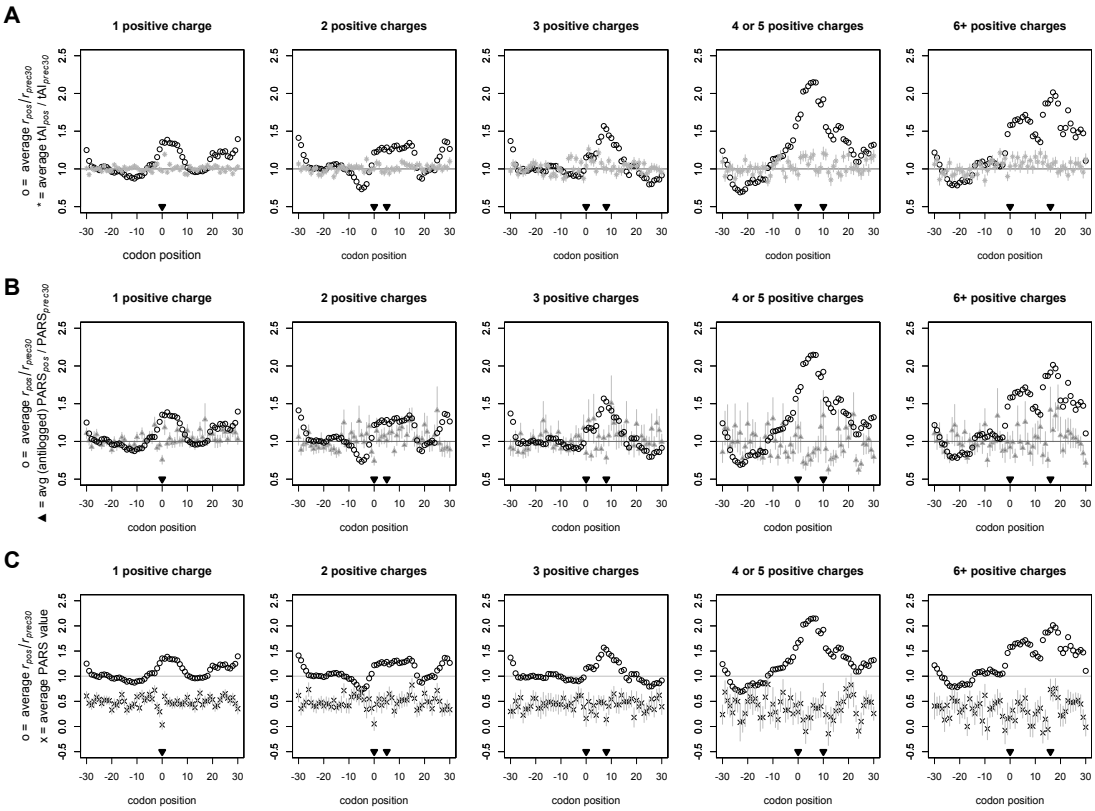

Figure S14. The effect of positive charge is not explained by covariance with codon usage or mRNA folding. In order to determine if global patterns of codon usage or mRNA secondary structure might in fact be contributing to patterns in ribosomal slowing we observe after clusters of positive charges, we also examined the relative changes in tAI and PARS values after the clusters. Within a given transcript, the relative increase or decrease in codon optimality at each position surrounding the charged cluster was calculated by dividing the measured ribosomal density at some codon position (tAI_{pos}) (i.e., at some position before/after the charged residue is added) by the average tAI of the thirty codons preceding the first coded-for charge in the cluster within that transcript (tAI_{prec30}). The mean relative change in tAI after a cluster positive charges was then calculated by aligning all transcripts with a given cluster size by the first charge in each cluster and calculating the average ratio (tAI_{pos}/tAI_{prec30}) in each codon site surrounding the cluster. We similarly calculated the relative increase or decrease in propensity for double-stranded structure, as quantified by PARS values, at each position surrounding the charged cluster. As PARS values as originally published [38] are logged ratios, we first took the antilog of all PARS values (making all of them positive) in order to be able to calculate relative increases or decreases in the values along transcripts by dividing the antilogged PARS value at some codon position surrounding the encoded charge cluster ($PARS_{pos}$) by the average PARS of the thirty codons (all previously antilogged) preceding the first coded-for charge in the cluster within that transcript ($PARS_{prec30}$). This method is conservative as taking the antilog will result in PARS values indicating single-strandedness being sandwiched between 0 and 1, but with PARS values indicating double-strandedness spread above 1. Hence increases in double-stranded propensity will be exaggerated. The average relative change in either tAI or PARS (mean tAI_{pos}/tAI_{prec30} or $PARS_{pos}/PARS_{prec30}$) at a given position after a cluster was then calculated by aligning all identified regions of a given cluster size according to the first charge present in each cluster and calculating the average ratio in positions increasingly distant from the first positive charge of the aligned clusters. Positive charges in a cluster may be coded for anywhere between the two downturned triangles. An average r_{pos}/r_{prec30} above one indicates a relative local increase in ribosomal density in that position across transcripts (as in Fig. 1). **A.** An average tAI_{pos}/tAI_{prec30} below one indicates the codons in that position across transcripts tend to decrease in optimality on average relative to the

average tAI of the preceding 30 codons across transcripts, while a ratio above one signifies an increase in optimality. We find that differential codon use in the vicinity of positive charges cannot explain the charge slowing effect. We observe no correlation between relative changes in ribosomal density and tAI after the first charge in the cluster ($0 \geq x \leq 30$ in this Figure, part A; Spearman P, left to right: 0.93, 0.73, 0.22, 0.17, 0.65). For a more relaxed test we then compared, for each plot in Fig. 5, the relative changes in codon optimality (tAI_{pos}/tAI_{prec30}) seen after the start of each cluster at x=0 until the point where relative change in ribosomal density (r_{pos}/r_{prec30}) drops back to previous levels ($y = 1$) to the tAI_{pos}/tAI_{prec30} values seen in all other surrounding plotted sites (i.e. those sites lacking charge-induced pausing). If anything, relatively more optimal ($tAI_{pos}/tAI_{prec30} > 1$) codons are coded for during periods of elevated ribosomal occupancy for clusters comprising 6 or more encoded cations, while no difference in optimality is detected in codon usage during elevated ribosomal occupancy compared to surrounding codon usage for other-sized charge clusters (Mann Whitney U-test P values, left to right in this Figure, part A: 0.96, 0.20, 0.07, 0.07, 0.003). Hence we conclude that changes in codon bias are not responsible for the slowing patterns associated with positively charged residues (Fig. 5), as expected if rare codons do not slow ribosomes (Fig 3A,B). **B.** An average relative change in (here antilogged, see Methods) PARS values (i.e. $PARS_{pos}/PARS_{prec30}$) plotted above one indicates a greater likelihood of double-stranded structure in that position on average relative to preceding sequence, while a ratio less than one indicates a decrease in propensity for double-strandedness relative to the preceding 30 codons. We find that the slowing effect of positive charge cannot be explained by mRNA folding in the vicinity of positive charges. There is no correlation between the relative change in PARS values ($PARS_{pos}/PARS_{prec30}$) after the first charge in the cluster (this Figure, part B, $0 \geq x \leq 30$) and relative changes in ribosomal density (Spearman P, left to right: 0.44, 0.68, 0.97, 0.99, 0.15), which we may have expected to observe if RNA structure has a local effect on ribosomal slowing. Likewise, under such a local-slowing hypothesis, we should expect to see a significant difference in the average PARS ratios seen amongst the sequence between $x = 0$ and the point at which elevated ribosomal density curve (r_{pos}/r_{prec30}) drops back to $y = 1$ versus PARS ratios in surrounding plotted sites. Such a difference, however, is seen only in the 2-charge plot (this Figure, part B; Mann Whitney U-test P values, left to right: 0.17, 0.0006, 0.24, 0.08, 0.60). If we instead assume that downstream structure has a pausing effect observable more upstream, a more appropriate test is to compare the PARS ratios from $-30 \geq x < 0$ to those from $0 \geq x \leq 30$. In this case we observe no significant difference in relative propensity for double-strandedness before or after positive charges apart from in the case of a single positive charge alone (this Figure, part B; Mann Whitney U-test, left to right: 0.004, 0.07, 0.12, 0.08 [with the mean $PARS_{pos}/PARS_{prec30}$ decreasing on average after the start of the cluster], 0.60). We note that this version of the test is exceedingly conservative as PARS values had to be antilogged before informative ratios could be calculated. This means that previously negative values (indicating single-strandedness) will now be sandwiched in between 0 and 1, while formerly positive values (indicating double-strandedness) now span a range of values above one. Hence normalizing the PARS score at a given position by the average PARS value of the preceding 30 codons will exaggerate not only the importance of structured versus free-form RNA, but will also exaggerate small differences in the magnitude of PARS values already denoting double-strandedness. **C.** An alternative calculation showing that RNA structure does not account for the pausing observed near positive charges. Note this Figure does not show the change in PARS values relative to the preceding sequence (as in B), but the average magnitude of the PARS value in that position across aligned transcripts. An average of PARS values plotted above zero indicates a greater likelihood of double-stranded structure in that position on average, while a mean value less than one indicates a propensity for single-strandedness. We find no correlation between the average PARS values after the first charge in the cluster ($0 \geq x \leq 30$) and relative changes in ribosomal density (this Figure, part C; Spearman P, left to right: 0.77, 0.95, 0.87, 0.34, 0.09), as we might have observed if RNA structure has a local effect on ribosomal slowing. Likewise, if structure causes local slowing, we should see a significant difference in the average PARS values between $x = 0$ and the point at which elevated ribosomal density curve (r_{pos}/r_{prec30}) drops back to $y = 1$ versus PARS values in surrounding plotted sites. We do not however detect such a difference (this Figure, part C; Mann Whitney U-test P values, left to right: 0.66, 0.17, 0.30, 0.27, 0.90). Examining whether downstream structure has a pausing effect observable further upstream, we

then compare the PARS ratios from $-30 \geq x < 0$ to those from $0 \geq x \leq 30$. In this case we observe no significant difference in relative propensity for double-strandedness before or after positive charges (this Figure, part C; Mann Whitney U-test, left to right: 0.98, 0.98, 0.97, 0.27, 0.90).

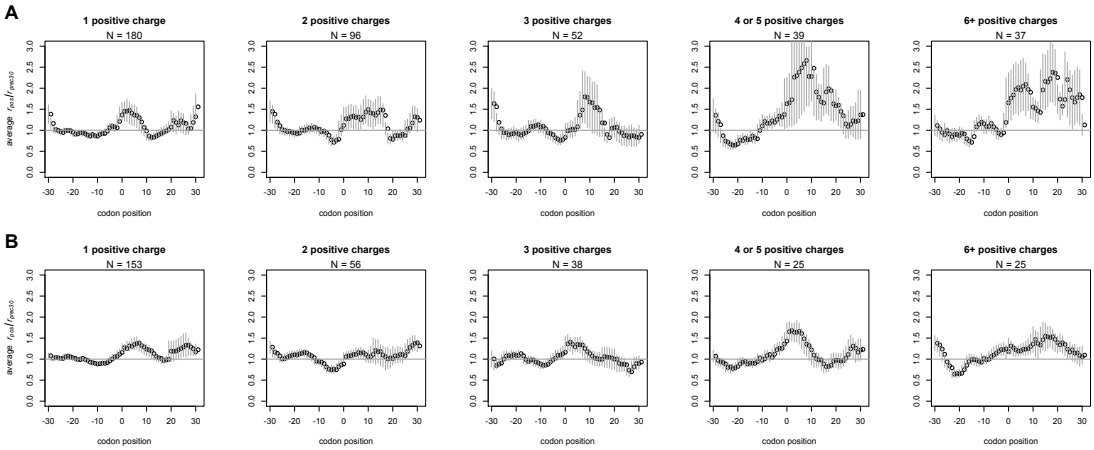

Figure S15. Genes with either high or low footprint coverage both produce consistent slowing patterns after positive charge clusters. To ensure that noise in the location of footprints among genes with fewer overall footprints is not an issue for analysis, we redrew our r_{pos}/r_{prec30} plots surrounding positive charge clusters using both the bottom half and top half of all genes according to their footprint saturation. Note that in this analysis we do not normalize the footprint counts per codon per gene by mRNA levels. This is because we are not interested in footprint coverage per transcript (as we might be if considering rates or mechanistic issues), but in the statistical power that the total footprint coverage per gene gives us, regardless of the number of transcripts that the footprints were captured from. Areas under the curve were measured as in the main text (see Figure 1). In each case we find similar results to those presented in the main analysis (Figure 5), namely that positive charges additively slow ribosomes. **A. Bottom half of genes**: Regression of *area under curve* ~ *cluster size*, slope = 4.8, r^2 = 0.79, P = 0.027. **B. Top half of genes**: Regression of *area under curve* ~ *cluster size*, slope = 0.96, r^2 = 0.74, P = 0.039.

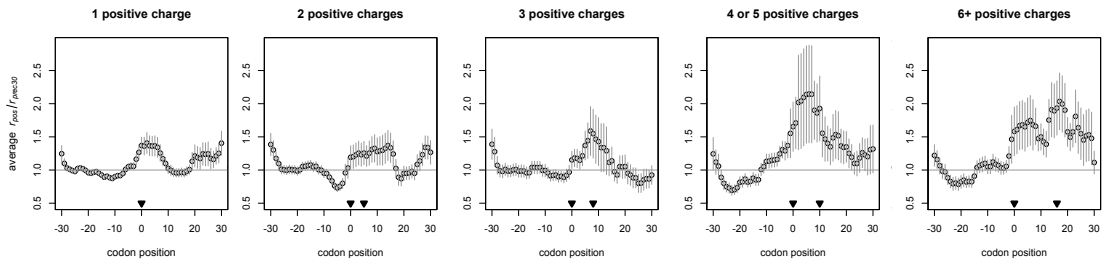

Figure S16. Figure 5 redone on the non-redundant footprint set. We wanted to confirm that the exclusion of footprints which map to two or more potential locations in the genome was not systematically biasing our estimates of ribosomal density. For this reason we replotted the average relative change in ribosomal density within a gene upon translation of encoded positive charge clusters using our non-redundant footprint set (see the end of the Methods section), in effect only considering those locations in the genome to which footprints uniquely map. Considering solely these regions in the transcriptome to which footprints can only ever be mapped unambiguously still shows positive charges additively slow translation.

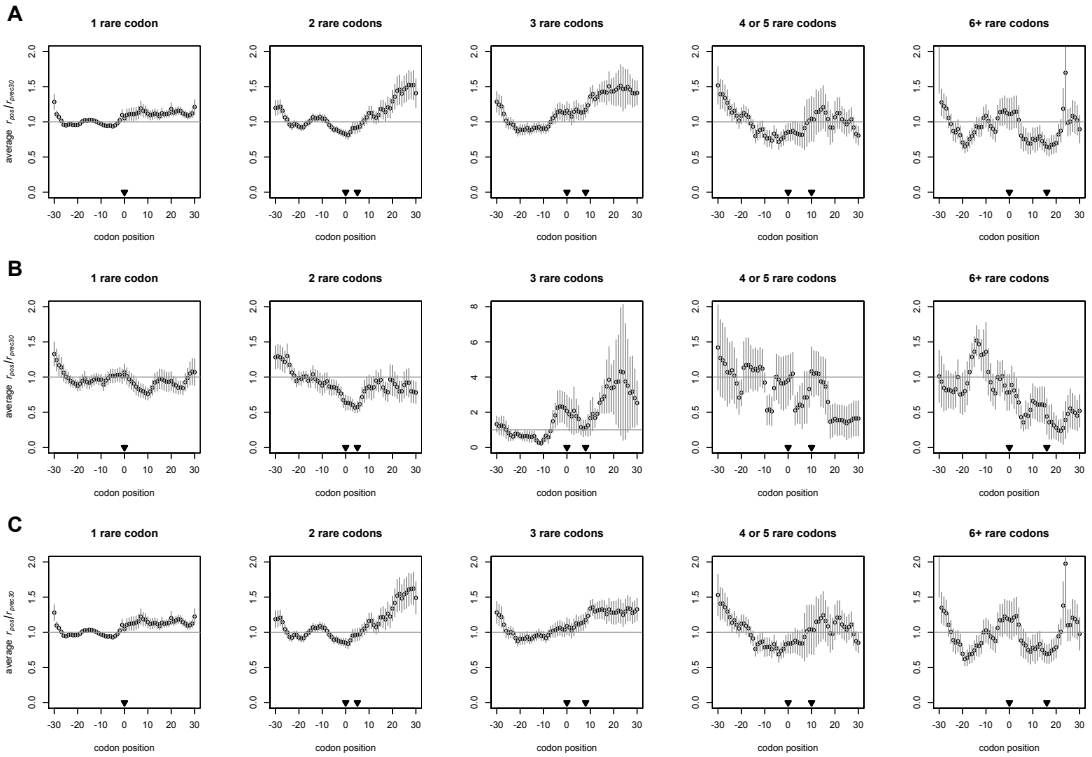

Figure S17. Figure 2 redone on the non-redundant footprint set. We wanted to confirm that the exclusion of footprints which map to two or more potential locations in the genome was not systematically biasing our estimates of ribosomal density. For this reason we replotted the average relative change in ribosomal density within a gene upon translation of rare codon clusters using our non-redundant footprint set (see the end of the Methods section), in effect only considering those locations in the genome to which footprints uniquely map. Considering solely these regions in the transcriptome to which footprints can only ever be mapped unambiguously still shows rare codons do not slow translation. **A)** All genes with rare codon clusters. **B)** Genes with rare codon clusters which have 0 or 1 positive charges coded for in the last 30 codon positions plotted. These plots represent the net effect of tAI on ribosomal density with the bulk of the effect of positive charge removed. **C)** Genes with rare codon clusters which have 2 or more positive charges in the last 30 codon positions plotted.

Note S1 is included in augmented form as Chapter III of this thesis.

Note S2

Only positive charge is capable of explaining the region of strongest translational pausing within transcripts

Having identified, within a given transcript, the two 10-codon windows with the largest difference in ribosomal densities, we then determined how often the denser region was more associated with each potentially slowing feature—positive charge, less optimal codons, pairs of rare codons, pairs of rare 6-mers, or a window of mRNA double-stranded structure immediately downstream (see Methods). If positive charge can indeed explain the greatest ribosomal deceleration within that transcript then across mRNAs we should expect that, even though subject to some stochasticity in ribosomal flow, the difference in the number of positive charges between the two windows (number of charges in high occupancy window-number of charges in low occupancy window) positively correlate with the difference in average ribosomal occupancy between the two windows (ribosomal occupancy in high occupancy window-ribosomal occupancy in low occupancy window) such that the excess magnitude of positive charge pairs with an excess magnitude of ribosomal density. Similarly if rare pairs are responsible for slowing, we should observe that the excess magnitude of rare pairs positively correlates with an excess magnitude of ribosomal density across transcripts, implying the more occluded window tends to contain more rare pairs. The same is true for rare 6-mers. However if basal codon optimality is responsible for the extremes in ribosomal occupancy between the two intra-transcript windows then we should expect a negative correlation between the difference in tAI between the two windows, meaning that across transcripts the window with the lower tAI tends to also be the window with more ribosomal footprints.

Of all the putative slowing features we consider, only charge is more often than not associated with the higher occupancy window within each transcript. Comparing each pair of intra-transcript windows across genes, we find that an excess of ribosomal density indeed correlates with an excess of positive charges as expected (Spearman rho 0.08, P = 6.4e-09). We are unable to detect a correlation between an excess of ribosomal density and an excess of rare codon pairs (Spearman P = 0.16), while that between the difference in density and an excess of rare 6-mers, while significant, is slight (Spearman rho = -0.04, P = 0.0066) and goes in the opposite direction expected were rare 6-mers capable of explaining slowing (considering only those genes which have at least one rare pair or rare 6-mer in either window, respectively). The correlation between difference in ribosomal density with difference in tAI between the two windows also goes in the wrong direction to explain pausing (Spearman rho 0.05, P = 0.00056), i.e. more occluded windows in fact tend to have higher (more optimal) tAIs. We do detect a negative correlation between the difference in the number of rare pairs between the two windows and the tAI of the rare pair (defined as the geometric mean of the tAIs of the two individual codons) (Spearman rho -0.28, P = 1.711e-05), implying that the pairs of rare codons in the higher occupancy window do indeed tend to be "less optimal". Nevertheless, the fact that the difference in number of a specific rare pair between windows and the corresponding mean difference in ribosomal density for windows containing that pair type negatively correlate (Spearman rho -0.15, P = 0.027) illustrates that more rare pairs, even if low in tAI, still more commonly associate with faster-travelling ribosomes.

Note S3

Trend of slowing increasing with charge is not random

To ask whether the trend we show in Figure 3C might result randomly given the specific genes we have used and their corresponding footprints, we performed the following test. We started with lists of all genes used in each positive charge plot in Figure 5. In the case of for example the 1-positive charge clusters, we counted the total number of identified 1-positive charge segments which went into making the 1-positive charge plot in Figure 5. For each iteration we identified the same number of segments (each time from a random gene in that list, and from a random location within that gene) and for each 'pseudo one-positive charge segment' we calculated r_{pos}/r_{prec30}; then after each iteration (out of 1000 total) we calculated the mean r_{pos}/r_{prec30} for that iteration. We then performed the similar randomizations for the other positive charge cluster sizes separately.

We partitioned the r_{pos}/r_{prec30} results at random into 1000 sets for which area under the curve plots can be calculated. For each set, we calculated the regression of $y \sim x$, where y is the vector of the area under the curves calculated from the randomization results (in exactly the same way as done in the main text, see Methods and Figure 1), and x is the vector of the average number of positive charges in each cluster used in the original analysis (1, 2, 3, 4.328, 6.823). Our randomization P value is then calculated as $P = (m+1)/(n+1)$, where m is the number of randomized sets for which the regression P value is significant and the slope is greater than or equal to that observed, and n is the sample size (1000). We find that the chance of detecting the trend we show in Figure 3C at random from just those genes used to make the Figure is indeed low (P = 0.011).

Table S1.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. rare codons score	1	354	345	338	232	8.7e-07 (2.6e-06)
	0	409	397	398	494	0.0015 (0.0045)
	-1	483	503	509	519	0.71
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	9.3e-06 (3.7e-05)	6.4e-08 (2.6e-07)	4.6e-09 (1.8e-08)	<2.2e-16 (8.8e-16)	-
B. rare pair score	1	60	73	49	25	2.5e-05 (7.5e-09)
	0	1074	1057	1084	1148	0.23
	-1	112	115	112	72	0.0063 (0.019)
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	9.0e-05 (0.00036)	0.0027 (0.011)	7.5e-07 (3.0e-06)	1.9e-06 (7.7e-06)	-

Table S1. Table 1 of the main text redone using rare codons which are defined to occur with genomic infrequency shows rare codons do not slow ribosomes. In the main text we investigate whether non-optimal codons, i.e. those with low tAI scores, might slow codons and find that they do not. To ensure that our finding that these 'rare' codons do not slow ribosomes does not simply hinge on our definition of 'rare', we have repeated the analysis using an alternative definition. Here, we define 'rare' codons according to their actual frequency in the genome as measured from our set of filtered genes. This rare set, of equal size to the rare tAI set, comprises the following codons: CGG, CGC, CGA, TGC, CCG, CTC, GGG, GCG, CGT, CCC, CAC, TGT, ACG, TCG, AGG. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. Rare (infrequent) codons and codon pairs tend to be found more in the less dense (faster translated) window. Similarly, the presence of rare pairs and rare codons decreases in the slowly-translated windows as the difference in degree of ribosomal slowing grows.

Table S2.

Protein expression quantiles	**q1** (least expressed)	**q2**	**q3**	**q4** (most expressed)
Number of genes with identified rare codon cluster	90	67	99	345

Table S2. **Genes with identified rare codon clusters are not disproportionately sampled from lowly expressed genes.** Could it be that large changes in ribosomal occupancy are not observed after rare clusters (Fig. 2A, Fig. 3A) because the clusters we identify are more likely to come from lowly expressed genes, i.e. genes which do not have high translation levels and for which it may be less likely that ribosomal footprints will be sampled? We used the average footprint count of a gene (total number of footprints within the coding sequence divided by gene length) as a proxy for protein expression levels. If anything, there are more genes with non-optimal codon clusters from genes which have more footprint reads (χ^2, $P < 2.2e\text{-}16$) so we do not consider this an issue.

Table S3.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. *Z* score	**1**	**483**	**517**	**479**	**522**	**0.39**
	0	**337**	**334**	**340**	**347**	**0.96**
	-1	**426**	**394**	**426**	**376**	**0.21**
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	**0.063**	**5.2e-05 (0.00021)**	**0.084**	**1.2e-06 (5.0e-06)**	**-**
B. *Z* score when charge score = 0	**1**	**94**	**99**	**84**	**87**	**0.68**
	0	**71**	**59**	**70**	**57**	**0.48**
	-1	**91**	**91**	**88**	**63**	**0.084**
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	**0.88**	**0.61**	**0.82**	**0.060**	**-**
C. *Z* score adjusted for amino acid usage	**1**	**323** *358*	**340** *340*	**314** *276*	**251** *265*	**0.0020 (0.0060)** *0.00013 (0.00039)*
	0	**373** *412*	**393** *424*	**381** *435*	**382** *436*	**0.91** *0.83*
	-1	**330** *256*	**285** *254*	**283** *267*	**267** *199*	**0.056** *0.0095 (0.028)*
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**0.81** *4.4e-05 (0.00018)*	**0.031 (0.12)** *0.00048 (0.0019)*	**0.22** *0.73*	**0.51** *0.0025 (0.010)*	**-**

Table S3. Sequence similarity to the yeast Kozak sequence cannot explain the greatest slowing within transcripts. Given that transcript similarity to the Shine-Dalgarno sequence has been shown to slow ribosomes in bacteria due to interactions of the sequence with components of the ribosomal RNA [17], we wondered whether translation speed in yeast might not be modulated by codon usage per se but by the ability of ribosomes to bind to transcript sequence which mirrors the eukaryotic Kozak sequence. Specifically, we wanted to determine whether codons which are in high-ribosomal occupancy windows within a gene might be more likely to correspond to the Kozak sequence (as compared to codons in low-occupancy windows within the same genes) and hence bind ribosomes, slowing translation. We first determined which codons were enriched in the Kozak sequence relative to the codon frequencies seen throughout the yeast genome at large using a simple randomization. Nucleotide frequencies at each position of the Kozak sequence in yeast were taken from Cavener and Ray 1991 [57]. To determine the frequencies of all the possible 'codons' among the Kozak sequence space, we randomly created 20000 possible Kozak sequences from the delineated nucleotide frequencies at each site in the consensus sequence. We then counted all possible triplet 'codons' within each sequence,

regardless of reading frame (since we assume that as the ribosome traverses RNA, it may bind the Kozak sequence regardless of the surrounding reading frame). The counts of all possible RNA triplets that we observe within our simulated sequences are the observed 'codons' within the Kozak sequence. In order to determine whether or not certain codons are over- or under-used in the Kozak sequence, we compare them to the counts of codons observed (again in any reading frame) across 20000 randomized sequences derived from the basal codon frequencies in the S. cerevisiae genome and of the same length as the Kozak sequence. We calculate *Z*, a measure of the over- or under-usage of a particular codon within the Kozak sequence (as compared to the rest of the genome) as Z_{codon} = [Observed codon count (in Kozak sequence) – Expected count (from genome frequencies)] / Expected SD of codon. We can then perform a test similar to the one in Methods V, but where we consider possible slowing codons to be those with a positive *Z* (GAT GAC AAC TGC CAA GGC GTA GTC TAT ACA TGG ATA CAT AAA TGT AAT ATG). A score of 1 indicates there are more codons with positive *Z* within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. A. Similarity to Kozak sequence cannot explain slowing in several quantiles (binomial tests), nor can it explain increased slowing (χ^2 tests). B. Even when the number of positive charges is the same between the two windows, we do not detect a significant contribution of similarity to Kozak sequence to slowing. C. Controlling for amino acid usage in two different ways, we detect no contribution of similarity to Kozak sequence to slowing; in fact, as the degree of slowing increases, the ability of Kozak similarity to explain slowing decreases (χ^2 tests). Method one (in bold): a gene is scored '1' if the slow window contains more codons with positive *Z*, '-1' if it contains fewer. Method two (in italics): the magnitude of all the positive *Z* values is averaged in each window, and the gene is scored '1' if the slower window has a higher average *Z*, '-1' if its average *Z* is lower.

Table S4.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value
A. aatAI score	**1**	**18** *634*	**20** *649*	**24** *637*	**28** *613*	**0.45** *0.79*
	0	**255** *5*	**220** *5*	**238** *2*	**279** *7*	**0.054** *0.57*
	-1	**24** *606*	**24** *590*	**15** *606*	**22** *626*	**0.46** *0.78*
	Binomial test on +1 and -1 charge score counts, P value	**0.44** *0.44*	**0.65** *0.10*	**0.20** *0.39*	**0.48** *0.73*	**-**
B. aatAI score (charge score = 0, original Δ*r* quantiles)	**1**	**1** *133*	**3** *129*	**2** *129*	**5** *97*	**0.44** *0.075*
	0	**42** *1*	**36** *0*	**31** *0*	**46** *1*	**0.34** *1*
	-1	**6** *124*	**4** *129*	**3** *107*	**1** *109*	**0.29** *0.39*
	Binomial test on +1 and -1 tAI score counts, P value	**0.125** *0.62*	**1.0** *1.0*	**1.0** *0.17*	**0.22** *0.44*	**-**
C. aatAI score (charge score = 0, recalculated Δ*r* quantiles)	**1**	**1** *123*	**3** *119*	**2** *137*	**5** *109*	**045** *0.35*
	0	**38** *0*	**33** *1*	**33** *0*	**51** *1*	**0.13** *1*
	-1	**4** *117*	**3** *119*	**5** *103*	**2** *130*	**0.81** *0.37*
	Binomial test on +1 and -1 rare pair score counts, P value	**0.38** *0.75*	**1.0** *1.0*	**0.45** *0.033*	**0.45** *0.20*	**-**

Table S4. Table 1 tAI score tests controlled for amino acid content. Could differences in amino acid usage between the two windows be biasing our result that neither codon usage nor rare pairs slow ribosomes (Table 1)? It could be that certain amino acids only have relatively high or low tAIs, and a preponderance of such amino acids in one window over the other could cause an apparent preference for (non-)optimal codons which is in fact a preference for a certain amino acid. For this reason we tested whether differences in amino acid usage between the high and low ribosomal occupancy windows within a transcript systematically alter the tAI scores (and hence the resulting interpretation of the contribution of codon usage to ribosomal density) in our window comparison analysis. To do this we identified the same high and low ribosomal occupancy windows within a transcript as above. This time, however, we considered only amino acids which are coded for at least once within each window. Within each intra-transcript window, we identified all codons that code for amino acid *x* and quantified the contribution of tAI to ribosomal occupancy using two approaches: **Method 1)** The average tAI of all the codons coding

for amino acid (aa) *x* was calculated for each window, and that amino acid was assigned an aa-tAIscore of 1, 0, or -1, depending on whether the tAI in the higher ribosomal occupancy window was lower (and hence capable of explaining the increased ribosomal density), the same, or higher than that in the other window, respectively. All of the aa-tAI scores for a given gene were counted independently—in other words, for a given gene it was possible to calculate more than one aa-tAI score, and all these aa-tAI scores contributed to the final matrix. **Method 2)** The average tAI of all the codons coding for amino acid *x* in each window was calculated, similarly to Method 1, but a tAI score is not yet assigned. Instead, the average tAI is first determined for each amino acid present in both windows, and then average tAIs (each the average for a particular amino acid) are themselves averaged to come up with a single aa-tAI for each window. Then, a single tAI score is assigned to that gene by comparing the average aa-tAIs in each window similarly to above. Bold = method 1, italic = method 2. Original Δr quantiles means the same quantile boundaries used in the main analysis were used, whereas recalculated Δr quantiles are drawn from only those genes considered in this amino-acid adjusted analysis. The P value for χ^2 tests with fewer than 5 observations in any square was calculated by resampling the observations without replacement and noting how many times (r) the χ^2 value of the resampled set was greater than or equal to the observed. P was then calculated as $(r+1)/(n+1)$, where n is the number of iterations performed (1000). **A.** Upon controlling for differential amino acid content in the two windows as detailed above, the result that tAI cannot explain patterns of slowing is still robust. Additionally we no longer detect a decrease in the ability of tAI to explain pausing in the upper quantiles as observed in Table 1A. **B** and **C** show the effect of tAI (adjusted for amino acid use) in only those pairs of intra-transcript windows which have the same number of positive charges between them.

Table S5.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. charge score	**1**	**552**	**569**	**587**	**610**	**0.36**
	0	**249**	**241**	**266**	**250**	**0.73**
	-1	**471**	**462**	**419**	**412**	**0.11**
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	**0.012 (0.048)**	**0.00095 (0.0038)**	**1.3e-07 (5.2e-07)**	**6.4e-10 (2.6e-09)**	**-**
B. tAI score	**1**	**632**	**615**	**607**	**577**	**0.46**
	0	**0**	**0**	**0**	**0**	**-**
	-1	**640**	**657**	**665**	**695**	**0.50**
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	**0.84**	**0.25**	**0.11**	**0.0010 (0.0040)**	**-**
C. rare pair score *rare 6-mer score*	**1**	**176** *267*	**175** *256*	**179** *234*	**116** *156*	**0.00066 (0.0020)** *3.3e-07 (9.9e-07)*
	0	**909** *712*	**875** *705*	**910** *730*	**1022** *773*	**0.0041 (0.012)** *0.28*
	-1	**187** *293*	**222** *311*	**183** *308*	**134** *343*	**7.7e-05 (0.00023)** *0.24*
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**0.60** *0.29*	**0.021 (0.082)** *0.023 (0.092)*	**0.87** *0.0017 (0.0068)*	**0.28** *< 2.2e-16 (8.8e-16)*	**-**
C. PARS score *conservative PARS score*	**1**	**103** *335*	**84** *297*	**79** *297*	**58** *300*	**0.0054 (0.022)** *0.34*
	0	**499** *0*	**492** *0*	**519** *0*	**543** *0*	**0.38** *-*
	-1	**107** *374*	**133** *412*	**111** *412*	**108** *409*	**0.27** *0.46*
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**0.84** *0.15*	**0.0011 (0.0044)** *1.8e-05 (7.1e-05)*	**0.024 (0.096)** *1.8e-05 (7.1e-05)*	**0.00013 (0.00051)** *4.8e-05 (1.6e-05)*	**-**

Table S5. Table 1 done again on the amino acid-starved footprint set [29]**.** Only positive charge is systematically capable of explaining ribosomal slowing, including the severest slowing**.** Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. A low codon optimality, if anything, tends to pair more with the less dense (faster translated) window.

Similarly, not only do rare pairs and rare 6-mers tend to be found more often in the faster-translated window, but their presence decreases as the difference in degree of ribosomal slowing grows. Additionally, a greater likelihood of transcript secondary structure at or just before the identified window is associated not with the more occluded windows, but with the less dense (faster translated) ones, and the presence of secondary structure in fact decreases as the difference in ribosomal slowing between the windows increases. Positive charge however is consistently associated with the higher density (more slowly-translated) window.

Table S6.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. hydropathy score	1	490	517	530	535	0.50
	0	215	191	216	214	0.56
	-1	541	537	499	496	0.34
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	0.12	0.56	0.35	0.24	-
B. polarity score	1	557	576	580	627	0.21
	0	206	168	158	179	0.065
	-1	483	501	507	439	0.12
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	0.024 (0.096)	0.024 (0.096)	0.029 (0.12)	9.4e-09 (3.7e-08)	-
C. negative charge score	1	547	539	507	584	0.14
	0	270	271	239	273	0.39
	-1	429	435	499	388	0.0024 (0.0072)
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	0.00018 (0.00072)	0.0010 (0.0040)	0.83	3.5e-10 (1.4e-9)	-
D. positive charge score	1	573	586	637	717	0.00014 (0.00043)
	0	258	259	236	207	0.0589 (0.18)
	-1	415	401	372	322	0.0038 (0.011)
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	5.6e-07 (2.2e-06)	4.3e-09 (1.7e-08)	<2.2e-16 (8.8e-16)	<2.2e-16 (8.8e-16)	-

Table S6. Positive charge best explains the slowest translated regions within transcripts compared to other physiochemical properties of amino acids. While we find that positive charges slow ribosomes, we wanted to control for the effects of other physiochemical properties of amino acids, specifically hydropathy (Phe, Val, Leu, Ile, Met), polarity (Asn, Gln, Ser, Thr, Cys, Tyr) and negative charge (Asp, Glu). These groups of amino acids, however, do not lend themselves to the r_{pos}/r_{prec30} analysis we carry out in the main text (See Figures 1-5) in the same way that positive charge does. The r_{pos}/r_{prec30} analysis is suited to positive charges because they

cluster in a way that gives us reasonable sample sizes given our constraints, i.e. the number of positive charges we require in the cluster and the additional requirement that there be no surrounding positive charges outside of the cluster. In the case of the other amino acid groups, there are either too many constituent members of the group and which are used too frequently (e.g. hydropathy) to define isolated 'clusters' for investigation, or the amino acids are used too rarely as clusters away from positive charges, and are of insufficient cluster sizes to establish any slowing trends (e.g. negative charges). We therefore compared the effects of these other physiochemical properties of amino acids by comparing the amino acids encoded by the highest-ribosomally occupied vs. lowest-occupied windows within genes. The analysis was carried out similarly to the way Table 1 was created in the main text, only this time counting different amino acids depending on the physiochemical property being investigated. We find that, on the whole, only positive charge can robustly explain the slowing patterns we observe. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. **A.** Hydrophobic residues (Phe, Val, Leu, Ile, Met) cannot explain increased slowing as the difference in translation speed between the two windows increases (χ^2 P = 0.98). Additionally the proportion of genes which pass the hydrophobicity test compared to failing it is only significant in the fourth quantile (q4) (binomial P = 0.023). **B.** Polar residues (Asn, Gln, Ser, Thr, Cys, Tyr) cannot explain increased slowing as the difference in translation speed between the two windows increases (χ^2 P = 0.21). Additionally the proportion of genes which pass the polarity test compared to failing it is only significant in the fourth quantile (q4) (binomial P = 3.7e-08). **C.** Negative charges (Asp, Glu) cannot explain increased slowing as the difference in translation speed between the two windows increases (χ^2 P = 0.14). Additionally the number of genes which pass or fail the negative charge score test in the third quantile (q3) is not significantly different (binomial P = 0.83). **D.** Positive charge score, from Table 1, is shown for purposes of comparison.

Table S7.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. hydropathy score when charge score = 0	1	101	107	103	104	0.98
	0	54	36	50	36	0.11
	-1	101	106	89	67	0.019 (0.056)
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	1	1	0.35	0.0057 (0.023)	-
B. hydropathy score when charge score = -1	1	199	207	191	186	0.73
	0	70	52	60	53	0.32
	-1	149	149	113	84	2.5e-05 (7.5e-05)
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	0.0085 (0.034)	0.0025 (0.010)	9.0e-06 (3.6e-05)	5.0e-10 (2.0e-09)	-
C. hydropathy score when charge score = 0 or -1	1	300	314	294	290	0.78
	0	124	88	110	89	0.031 (0.093)
	-1	250	255	202	151	3.1e-07 (9.3e-07)
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	0.037 (0.15)	0.015 (0.060)	4.2e-05 (0.00017)	3.5e-11 (1.4e-10)	-
D. charge score when hydropathy score = 0	1	91	103	106	125	0.13
	0	54	36	50	36	0.11
	-1	70	52	60	53	0.32
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	0.11	5.1e-05 (0.00020)	0.00044 (0.0018)	6.9e-08 (2.8e-07)	-
E.	1	291	282	297	345	0.049 (0.15)

charge score when hydropathy score = -1						
	0	**101**	**106**	**89**	**67**	**0.019 (0.057)**
	-1	**149**	**149**	**113**	**84**	**2.5e-05 (7.5e-05)**
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	**1.2e-11 (4.8e-11)**	**1.5e-10 (6.0e-10)**	**2.2e-16 (8.8e-16)**	**2.2e-16 (8.8e-16)**	**-**
F. charge score when hydropathy score = 0 or -1	**1**	**382**	**385**	**403**	**470**	**0.0063 (0.019)**
	0	**155**	**142**	**139**	**103**	**0.011 (0.033)**
	-1	**219**	**201**	**173**	**137**	**0.00010 (0.00030)**
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**3.0e-11 (1.2e-10)**	**2.5e-14 (1.0e-13)**	**2.2e-16 (8.8e-16)**	**2.2e-16 (8.8e-16)**	**-**

Table S7. Positive charge explains slowing better than amino acid hydrophobicity. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. **A – C.** In those genes which fail the positive charge test (charge score = 0 or -1), we find that hydrophobicity cannot explain the increased slowing in these windows either (this table, χ^2 tests). For this reason we consider that while amino acids with hydrophobic side chains may be used more often in the vicinity of positive charge (this table, binomial tests), perhaps for certain structural motifs or because of the types of genes under consideration, they cannot responsible for the major slowing effect. **D – F.** Positive charge can explain the slowing in genes where hydrophobicity cannot.

Table S8.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. polarity score when positive charge score = 0	1	112	114	115	107	0.95
	0	51	39	43	35	0.34
	-1	93	96	84	65	0.074
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	0.21	0.24	0.033 (0.132)	0.0017 (0.0068)	-
B. polarity score when positive charge score = -1	1	253	246	216	211	0.12
	0	71	58	45	33	0.0014 (0.0042)
	-1	94	104	103	79	0.24
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	2.2e-16 (8.8e-16)	2.1e-14 (8.5e-14)	2.4e-10 (9.4e-10)	4.8e-15 (1.9e-14)	-
C. polarity score when positive charge score = 0 or -1	1	365	360	331	318	0.21
	0	122	97	88	68	0.0011 (0.0033)
	-1	187	200	187	144	0.019 (0.057)
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	3.0e-14 (1.2e-13)	1.3e-11 (5.4e-11)	2.6e-10 (1.0e-09)	3.8e-16 (1.5e-15)	-
D. positive charge score when polarity score = 0	1	84	71	70	111	0.0046 (0.014)
	0	51	39	43	35	0.34
	-1	71	58	45	33	0.0014 (0.0041)
	Binomial test on +1 and -1 charge score counts, P value	0.34	0.29	0.025 (0.1)	4.6e-11 (1.8e-10)	-

	(Bonferroni correction)					
E. positive charge score when polarity score = -1	1	296	301	320	295	0.79
	0	93	96	84	65	0.074
	-1	94	104	103	79	0.24
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	2.2e-16 (8.8e-16)	2.2e-16 (8.8e-16)	2.2e-16 (8.8e-16)	2.2e-16 (8.8e-16)	-
F. positive charge score when polarity score = 0 or -1	1	380	372	390	406	0.65
	0	144	135	127	100	0.036 (0.11)
	-1	165	162	148	112	0.0071 (0.021)
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	2.2e-16 (8.8e-16)	2.2e-16 (8.8e-16)	2.2e-16 (8.8e-16)	2.2e-16 (8.8e-16)	-

Table S8. Positive charge explains slowing better than amino acid polarity. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. **A – C.** In those genes which fail the positive charge test (charge score = 0 or -1), we find that polarity cannot explain the increased slowing in these windows either (this table, χ^2 tests). For this reason we consider that while amino acids with polar side chains may be used more often in the vicinity of positive charge (this table, binomial tests), perhaps for certain structural motifs or because of the types of genes under consideration, they cannot responsible for the major slowing effect. **D – F.** Positive charge can explain the slowing in some genes where polarity cannot.

Table S9.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. negative charge score when positive charge score = 0	1	106	114	105	100	0.81
	0	52	52	49	42	0.71
	-1	98	83	88	65	0.076
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	0.62	0.032 (0.13)	0.25	0.0079 (0.032)	-
B. negative charge score when positive charge score = -1	1	213	208	176	176	0.10
	0	89	76	64	66	0.15
	-1	116	124	124	81	0.010 (0.030)
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	9.8e-08 (3.9e-07)	4.7e-06 (1.9e-05)	0.0032 (0.013)	3.0e-09 (1.2e-08)	-
C. negative charge score when positive charge score = 0 or -1	1	319	322	281	276	0.11
	0	141	128	113	108	0.14
	-1	214	207	212	146	0.0010 (0.0030)
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	6.2e-06 (2.5e-05)	6.5e-07 (2.6e-06)	0.0022 (0.0088)	2.5e-10 (9.9e-10)	-
D. positive charge score when negative charge score = 0	1	129	143	126	165	0.081
	0	52	52	49	42	0.71
	-1	89	76	64	66	0.15
	Binomial test	0.0081	7.0e-06	8.1e-06	5.8e-11	-

	on +1 and -1 charge score counts, P value (Bonferroni correction)	**(0.032)**	**(2.8e-05)**	**(3.2e-05)**	**(2.3e-10)**	
E. positive charge score when negative charge score = -1	**1**	**215**	**228**	**287**	**242**	**0.0070 (0.021)**
	0	**98**	**83**	**88**	**65**	**0.076**
	-1	**116**	**124**	**124**	**81**	**0.010 (0.030)**
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	**5.8e-08 (2.3e-07)**	**3.2e-08 (1.3e-07)**	**5.5e-16 (2.2e-15)**	**2.2e-16 (8.8e-16)**	**-**
F. positive charge score when negative charge score = 0 or -1	**1**	**344**	**371**	**413**	**407**	**0.042 (0.13)**
	0	**150**	**135**	**137**	**107**	**0.059**
	-1	**205**	**200**	**188**	**147**	**0.011 (0.033)**
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**3.2e-09 (1.3e-08)**	**7.6e-13 (3.1e-12)**	**2.2e-16 (8.8e-16)**	**2.2e-16 (8.8e-16)**	**-**

Table S9. Positive charge explains slowing better than negative charge. Quantiles of the difference in average ribosomal density between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. **A – C.** In those genes which fail the positive charge test (charge score = 0 or -1), we find that negatively charged amino acids cannot explain the increased slowing in these windows either (this table, χ^2 tests). For this reason we consider that while amino acids with negatively charged side chains may be used more often in the vicinity of positive charge (this table, binomial tests), perhaps for certain structural motifs or because of the types of genes under consideration, they cannot responsible for the major slowing effect. **D – F.** Positive charge can explain the slowing in genes where negative charge cannot.

Table S10.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A.	**charge score 1 tAI score 1**	271	272	281	284	0.93
	charge score 1 tAI score -1	302	314	356	433	1.4e-06 (2.8e-06)
	Binomial test P value (Bonferroni correction)	0.21	0.090	0.0033 (0.013)	2.9e-08 (1.2e-07)	-
B.	**charge score 0 tAI score 1**	116	130	111	101	0.28 (0.56)
	charge score 0 tAI score -1	142	129	125	106	0.15 (0.30)
	Binomial test P value (Bonferroni correction)	0.12	1.0	0.40	0.78	-
C.	**charge score -1 tAI score 1**	203	195	171	140	0.0036 (0.0072)
	charge score -1 tAI score -1	212	206	201	182	0.47
	Binomial test P value (Bonferroni correction)	0.69	0.62	0.13	0.022 (0.09)	-

Table S10. The relationship of charge score to tAI score. Quantiles of the difference in average ribosomal occlusion between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. **A.** The ability of charge to explain slowing (charge score of 1) cannot be explained by concomitant use of suboptimal codons. A charge score of 1 more commonly pairs with a tAI score which cannot explain slowing (tAI score of -1), and increasingly so as the difference in ribosomal speeds between the two windows grows. **B**. These tAI scores are drawn from transcripts for which both intra-transcript windows have the same number of charges (charge score = 0) and hence such comparisons should be controlled for the effect of positive charge on ribosomal speed. Different tAI scores are equally distributed among quantiles, indicating the inability of tAI to predict either ribosomal slowing or the degree of ribosomal slowing even in the absence of an effect of charge on ribosomal speed. **C**. tAI does not systematically account for slowing in windows for which increased charge pairs with the faster window.

Table S11.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A.	**charge score 1 rare pair score 1**	70	81	64	48	0.034 (0.068)
	charge score 1 rare pair score -1	92	89	98	79	0.55
	Binomial test P value (Bonferroni correction)	0.10	0.59	0.0093 (0.03)	0.0075 (0.03)	
B.	**charge score 0 rare pair score 1**	34	36	31	14	0.015 (0.030)
	charge score 0 rare pair score -1	46	41	42	20	0.012 (0.024)
	Binomial test P value (Bonferroni correction)	0.22	0.65	0.24	0.39	
C.	**charge score -1 rare pair score 1**	71	62	49	24	2.1e-05 (4.2e-05)
	charge score -1 rare pair score -1	75	52	56	24	1.2e-05 (2.4e-05)
	Binomial test P value (Bonferroni correction)	0.80	0.40	0.56	1.0	

Table S11. The relationship of rare pair score to charge score. Quantiles of the difference in average ribosomal occlusion between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts. **A.** The ability of charge to explain slowing (charge score of 1) cannot be explained by concomitant use of rare pairs. A charge score of 1, if anything, tends to pair with a rare pair score which cannot explain slowing (rare pair score of -1). **B**. These rare pair scores are drawn from transcripts for which both intra-transcript windows have the same number of charges (charge score = 0) and hence such comparisons should be controlled for the effect of positive charge on ribosomal speed. Different rare pair scores are equally distributed among quantiles, indicating the inability of rare pairs to predict ribosomal slowing. Additionally, as the difference in the degree of ribosomal slowing increases (i.e. moving from q1 to q4), the number of rare pairs found in the higher occupancy window decreases (χ^2 test), demonstrating rare pairs cannot predict the magnitude of slowing even in the absence of an effect of charge on ribosomal speed. **C**. Rare pairs do not systematically account for slowing in windows for which increased charge pairs with the faster window.

Table S12.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A.	**charge score 1 PARS score 1**	40	32	48	29	0.12
	charge score 1 PARS score -1	78	67	82	73	0.64
	Binomial test P value (Bonferroni correction)	0.00060 (0.0024)	0.00056 (0.0022)	0.0036 (0.014)	1.6e-05 (6.4e-05)	-
B.	**charge score 0 PARS score 1**	24	13	12	11	0.062
	charge score 0 PARS score -1	86	77	84	61	0.17
	Binomial test P value (Bonferroni correction)	2.2e-09 (8.9e-09)	3.2e-12 (1.3e-11)	1.8e-14 (7.3e-14)	1.5e-09 (6.2e-09)	-
C.	**charge score -1 PARS score 1**	22	27	21	15	0.33
	charge score -1 PARS score -1	44	37	21	23	0.008 (0.016)
	Binomial test P value (Bonferroni correction)	0.0092 (0.037)	0.26	1	0.26	-
D.	**charge score 1 conservative PARS score 1**	140	126	163	155	0.14
	charge score 1 conservative PARS score -1	200	248	233	246	0.095
	Binomial test P value (Bonferroni correction)	0.0013 (0.0052)	2.8e-10 (1.1e-09)	0.00051 (0.0020)	6.4e-06 (2.6e-05)	-
E.	**charge score 0 conservative PARS score 1**	72	52	52	54	0.18
	charge score 0 conservative PARS score -1	86	77	84	61	0.17
	Binomial test P value (Bonferroni correction)	0.30	0.034 (0.14)	0.0076 (0.030)	0.58	-
F.	**charge score -1 conservative PARS score 1**	90	94	75	85	0.50
	charge score -1 conservative PARS score -1	121	111	101	108	0.60
	Binomial test P value (Bonferroni correction)	0.039 (0.16)	0.26	0.060	0.11	-

Table S12. The relationship of PARS score (double strandedness) to charge score. Quantiles of the difference in average ribosomal occlusion between the two windows identified within a transcript are shown, with q1 representing the smallest differences and q4 the largest. A score of 1 indicates the putative retarding feature is more present within the more occluded intra-transcript window; -1, less present; 0, present in both windows in equal amounts.

Table S13.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. charge score	**1**	**578**	**630**	**700**	**766**	**1.3e-06 (3.8e-06)**
	0	**270**	**256**	**255**	**242**	**0.67**
	-1	**496**	**457**	**388**	**335**	**5.2e-08 (1.6e-07)**
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	**0.013 (0.052)**	**1.7e-07 (6.8e-07)**	**< 2.2e-16 (8.8e-16)**	**< 2.2e-16 (8.8e-16)**	**-**
B. tAI score	**1**	**576**	**600**	**601**	**551**	**0.41**
	0	**0**	**0**	**0**	**0**	**-**
	-1	**768**	**743**	**742**	**792**	**0.53**
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	**1.8e-07 (7.2e-07)**	**0.00011 (0.00042)**	**0.00013 (0.00053)**	**5.2e-11 (2.1e-10)**	**-**
C. rare pair score *rare 6-mer score*	**1**	**180** *257*	**184** *205*	**160** *177*	**92** *115*	**9.9e-08 (3.0e-07)** *4.8e-12 (1.4e-11)*
	0	**904** *712*	**907** *711*	**958** *730*	**1101** *868*	**7.6e-06 (2.3e-05)** *4.6e-05 (0.00014)*
	-1	**260** *375*	**252** *427*	**225** *436*	**150** *360*	**2.0e-07 (6.0e-07)** *0.014 (0.042)*
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**0.00016 (0.00064)** *3.1e-06 (1.2e-05)*	**0.0013 (0.0052)** *<2.2e-16 (8.8e-16)*	**0.0011 (0.0044)** *<2.2e-16 (8.8e-16)*	**0.00023 (0.00092)** *<2.2e-16 (8.8e-16)*	**-**
C. PARS score *conservative PARS score*	**1**	**88** *301*	**88** *269*	**88** *289*	**59** *304*	**0.05** *0.46*
	0	**491** *0*	**510** *0*	**516** *0*	**554** *0*	**0.26** *-*
	-1	**155** *433*	**136** *465*	**130** *445*	**121** *430*	**0.21** *0.64*
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**2.1e-05 (8.2e-05)** *1.2e-06 (5.0e-05)*	**0.0016 (0.0065)** *4.5e-13 (1.8e-12)*	**0.0054 (0.022)** *9.4e-09 (4.4e-06)*	**4.4e-06 (1.8e-05)** *3.8e-06 (1.5e-05)*	**-**

Table S13. Table 1 done again, allowing the lower-occupancy window to have a ribosomal occupancy of 0.

Table S14.

		$q1_{\Delta r}$ (count)	$q2_{\Delta r}$	$q3_{\Delta r}$	$q4_{\Delta r}$	χ^2 test P value (Bonferroni correction)
A. charge score	**1**	**571**	**591**	**637**	**708**	**0.00051 (0.0015)**
	0	**267**	**241**	**256**	**213**	**0.08**
	-1	**419**	**425**	**363**	**336**	**0.0022 (0.0065)**
	Binomial test on +1 and -1 charge score counts, P value (Bonferroni correction)	**1.5e-06 (6.0e-06)**	**2.1e-07 (8.4e-07)**	**< 2.2e-16 (8.8e-16)**	**< 2.2e-16 (8.8e-16)**	**-**
B. tAI score	**1**	**595**	**589**	**573**	**539**	**0.35**
	0	**0**	**0**	**0**	**0**	**-**
	-1	**662**	**668**	**683**	**718**	**0.45**
	Binomial test on +1 and -1 tAI score counts, P value (Bonferroni correction)	**0.06 (0.24)**	**0.03 (0.12)**	**0.0021 (8.4e-03)**	**4.9e-07 (2.0e-06)**	**-**
C. rare pair score *rare 6-mer score*	**1**	**181** *126*	**162** *107*	**152** *85*	**87** *46*	**1.7e-07 (5.0e-07)** *1.8e-08 (5.4e-08)*
	0	**858** *382*	**912** *401*	**913** *425*	**1048** *499*	**1.0e-06 (3.0e-04)** *0.00035 (1.1e-03)*
	-1	**218** *199*	**183** *198*	**191** *196*	**122** *161*	**4.4e-06 (1.3e-05)** *0.15*
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**0.07** *6.1e-05 (2.4e-04)*	**0.28** *2.1e-07 (8.4e-07)*	**0.040 (0.16)** *3.0e-11 (1.2e-10)*	**0.018 (0.072)** *3.8e-16 (1.5e-15)*	**-**
C. PARS score *conservative PARS score*	**1**	**86** *301*	**72** *271*	**80** *290*	**57** *292*	**0.093** *0.065*
	0	**466** *0*	**509** *0*	**500** *0*	**543** *0*	**0.11** *-*
	-1	**154** *405*	**125** *435*	**126** *416*	**107** *415*	**0.032 (0.096)** *0.77*
	Binomial test on +1 and -1 rare pair score counts, P value (Bonferroni correction)	**1.3e-05 (5.2e-05)** *0.00010 (0.00040)*	**0.00020 (0.00080)** *7.2e-10 (2.9e-09)*	**0.0017 (0.0068)** *2.4e-06 (9.6e-06)*	**0.00012 (0.00048)** *4.2e-06 (1.7e-05)*	**-**

Table S14. Table 1 done again on the non-redundant footprint location set.

III. *Codon usage and translation rates: how can codon usage not predict ribosome occupancy but be commonly assumed to be associated with faster translation?*

Catherine A. Charneski & Laurence D. Hurst

Based on Note S1 from *PLoS Biol* (2013) 11(3): e1001508

Ever since suggested by Ames and Hartman (1963) there has been a growing thread throughout the literature that a codon specifying a rare tRNA might stall ribosomes as the enzyme awaits for the tRNA to enter its A-site. Indeed in some instances, it is now often simply presumed that different codons must affect ribosome velocity. For example, Higgs and Ran (Higgs and Ran 2008) assert that "it is differences in rates [of translation] between alternative cognate codons that are relevant for codon bias. The fact that codon bias occurs in a large number of bacterial genomes means that these rates must indeed differ". While this assertion discounts the possibility that translational accuracy might be important, nonetheless, we can ask how it is that there exists apparent, often-cited evidence for codon usage altering translation rates but at the same time we see no evidence that codon usage predicts ribosome occupancy in mouse embryonic stem cells (Ingolia, Lareau, Weissman 2011), *E. coli* (Li et al. 2012), or yeast (Qian et al. 2012; Charneski and Hurst 2013).

There are, we suggest two classes of explanation. First, we note that much evidence is indirect and/or fails to address alternative explanations. Second, and perhaps more interestingly, we argue that it is possible that under normal conditions codon usage should not predict translation rates, but out of normal (equilibrium) conditions codon bias may have a profound effect. This we suggest can explain apparently contradictory evidence.

1. Do experimental results support the view that codon usage modulates translational velocity: problems with predictions and covariates

One possibility is that the codon usage – translation rate assumption is so profoundly held that evidence for the effect is over-interpreted. Indeed, some studies cited as support that codon usage can influence translation rates offer only circumstantial evidence. In such studies the results are possibly consistent with a proposed role for codon usage in modulating ribosomal velocity, but this is not explicitly shown. For example Konigsberg and Godson (1983) showed differential codon usage between *dnaG* and a handful of other *E. coli* regulatory genes in comparison with 25 non-regulatory genes and speculated that codon usage may cause differential expression levels of these categories of protein. Another early study aligned the coding sequences of two genes with different translational pause sites (as indicated by gel electrophoresis), and upon noting differences in codon usage then claimed "we have now demonstrated that this particular codon usage… is directly responsible for the variable rate of polypeptide chain elongation observed" (Morlon et al. 1983). A more recent study (Boycheva et al. 2003) searched for codon pairs which were overrepresented in lowly expressed and highly expressed genes and denoted them hypothetical attenuating and non-attenuating pairs, respectively, but the authors were unable to

use experimental data to further investigate whether such codon pairs might in fact have any effect on ribosomal speed.

Similarly, quite a few experimental studies presuming to show influence of codon bias on gene expression assume that a lack or reduction in protein product reflects slower translation rates (Sørensen et al. 1989; Goldman et al. 1995; Cannarozzi et al. 2010; Nakamura and Sugiura 2011). However, a heterologous gene transfected into an organism with a different codon usage bias, or a gene engineered to code for the same protein product as the original but with less optimal codons might have unstable mRNA (Stanssens et al. 1986; Hoekema et al. 1987; Petersen 1987) or protein products (Kurland 1991). For instance Coleman et al. (2008) claim that substitutions of underrepresented codon pairs into poliovirus coding sequence cause decreased translation rates when such mRNAs are expressed in HeLa cells. Slowing of translation, however, is not explicitly shown and it is possible that the assays—reduced infectivity or reduced enzymatic activity—could be the result of e.g. structural errors in the proteins which are unrelated to translation speed.

Perhaps more crucially, it is far from clear that changing codon usage bias should greatly change protein titres if translation rate mediated by tRNA abundance is the sole force. After a ribosome has finished processing a transcript for a given gene, the chance that the freed up ribosome will then process a transcript from the same gene is low. The major effect of changing translation speed should thus be changes in cell growth rates not necessarily changes in titre of the protein concerned. More exactly, it has been suggested that changing the translation rate of an mRNA is only likely to directly influence the amount of the focal protein produced if that mRNA can capture a majority of all ribosomes in the cell (Andersson and Kurland 1990). In other words, faster translation of an mRNA is not likely to affect the resulting amount of the focal protein in that cell if there is no ribosome readily available to immediately start translating another copy of the same mRNA (provided translation initiation features allow prompt re-initiation). This proposition is supported by transgene studies by Kudla et al. (Kudla et al. 2009) who showed no correspondence between codon usage bias of upregulated versions of GFP, differing only at synonymous sites, and protein titre.

This group also revealed the importance of controlling for translation initiation features at the 5' end of a transcript. A number of other studies (Hoekema et al. 1987; Goldman et al. 1995; Irwin et al. 1995; Cannarozzi et al. 2010) have examined the effect of synonymous mutations including those near the beginning of transcripts where it is known mRNA structure can influence the frequency of translation initiation. For example Irwin et al. (1995) focused on the effect of substitutions of codon pairs at the 5' end, assuming that if they were translated slowly it would

prevent re-initiation of *lacZ* by another ribosome and hence reduce the amount of beta-galactosidase activity observable. But their test may have rather interfered with mRNA secondary structure important for translation initiation. It should also be noted that Irwin et al. (1995) found that *over*-represented rare pairs of codons in E coli could attenuate ribosomes. Their claim however was later disputed by Cheng and Goldman (2001) who could not confirm Irwin's findings. In another case, Goldman et al. (1995) observed that insertion of 9 consecutive low-usage leucine codons near the 5' end of a transcript blocked translation, but no similar effects were seen when the 9 consecutive codons were introduced further downstream, suggesting their results may be due to interference with 5' transcript folding required for initiation.

Even when the experiments are robust, further problems with covariates (i.e. alternative explanations) abound. For example the slowing (as inferred by reduced expression level) thought to be due to consecutive rare AGG or AGA codons in *E. coli* (Robinson et al. 1984; Misra and Reeves 1985; Varenne and Lazdunski 1986; Bonekamp and Jensen 1988; Spanjaard et al. 1990; Sørensen and Pedersen 1991; Rosenberg et al. 1993; Wang et al. 1994; Hu et al. 1996) may be due to tandem codons resembling the Shine Dalgarno sequence and interacting with the translating ribosome (Ivanov et al. 1992; Wen et al. 2008; Li et al. 2012) or indeed in some cases the positive charge on the incorporated arginine residue may slow the ribosome (Lu and Deutsch 2008; Charneski and Hurst 2013). Similarly, Sørensen et al. (1989) inferred average translation rates from the time required for *E. coli* to incorporate radioactive methionine into β-galactosidase containing inserts full of either rare or common codons. Their rare-codon insert, however, contained more (and more clustered) codons encoding positive charges, which may account for the slowing of ribosomes during translation of rare codons that they infer. A perfect test is indeed very hard to envisage as any change to codon usage is likely also to affect many aspects of the processing of the RNA, not least the RNA structure, for which we found some evidence of an (albeit marginal) effect on translation rates (Charneski and Hurst 2013).

While claims that changing the codon usage modulates levels of that protein of the modified gene because of changes in translation rate owing to tRNA availability should be treated with considerable caution, some more direct reports lack robust statistics. As an example, Varenne et al. (1984) report that translation rate along mRNA varies with tRNA availability at different codons (although some of the slowing they observe they say cannot be attributed to differential codon usage). They compared the distribution of electrophoretic intermediates to that predicted by assuming that tRNA concentration is the rate-limiting factor in ribosomal translocation, with the aim of investigating how well the prediction matched the observed. However, there are a few problems with the approach. 'Analogous peaks' between the observed and predicted were detected not by a stringent methodology but by attempting to locate matching peaks between

noisy curves by eye. Nor was it determined if detected slowing along the ribosome was significant. Additionally although a good correlation between the observed and predicted curves was claimed, a statistical test of a correlation, or any type of statistical test to establish similarity between the two curves, was not performed. We do not wish to assert that the conclusions of Varenne et al. are incorrect, just that they lack normal statistical support. Another group, examining the same ribosomal footprinting dataset as we do in Chapter II, found a correlation of codon usage bias with ribosomal density at 5' transcript ends and concluded that codon usage is responsible for the excess density observed at transcript starts (Tuller et al. 2010; Tuller et al. 2011). Aside from questions of whether this excess 5' density is an artefact of the footprinting method used (previously discussed in Ingolia et al. 2011; Charneski and Hurst 2013), we note that the statistical robustness of these analyses is lacking for three main reasons. First, footprints were allowed to map to more than one genomic location, which almost certainly artificially inflates and skews the types of sequences occluded by ribosomal footprints. Second, the correlation was performed on mean values calculated from aligned transcripts, not on a within-transcript basis, allowing for the possibility of emergent patterns which do not reflect an underlying mechanism. Third, neighboring codons were allowed to be occluded by the flanking (i.e. not active site) regions of footprints corresponding to ribosomes translating not that codon but neighboring codons, which means the correlations were done using non-independent data points and are thus statistically invalid.

2. Normal and abnormal conditions and the balance model of codon usage

Above we have suggested that the tendency to suppose a direct link, mediated by codon usage and tRNA abundances, has often led to alternative interpretations not being considered. While the problem of alternative explanations must always be an issue, we don't wish to suppose that there is no evidence that changes in codon usage bias cannot *sometimes* affect translation rates. We note, however, that the best of the evidence finds support for an effect under abnormal conditions. For example, rare Arg codons (AGG) can limit protein synthesis in *E. coli* compared to the same amino acid sequence comprised of non-rare codons (CGT) (Robinson et al. 1984). This effect, however, was only observed under extreme conditions involving multiple consecutive rare codons and transcription at very high levels. Such runs of consecutive rare codons in highly expressed transcripts are unlikely to be observed in endogenous genes.

Similarly, Pedersen (1984) found a less than two-fold difference in the translation rate between rare and common codons upon comparing the speed of translation of ribosomal proteins, but only when they were expressed in high-copy plasmids with an up-promoter mutation, presumably

increasing drain on the tRNA pool. Similarly Misra and Reeves (1985) reported stalling (as inferred by accumulation of an intermediate peptide) at a rare Arg codon which could be rescued by providing $tRNA_{Arg}$(AGA), but this stalling effect was observed upon transcribing the gene from a multicopy plasmid and may not reflect *in vivo* conditions. Komar et al. (1999) showed that substitution of 16 rare for frequent synonymous codons in a 21-codon stretch resulted in loss of a protein intermediate as visualized by gel electrophoresis, but the cloned transgene was expressed from the high-expression viral T7 promoter. Kudla et al.'s (Kudla et al. 2009) demonstration that codon usage bias predicts growth rates is also consistent with an effect on translation rates for grossly upregulated genes. Most recently, an experimental techinique whereby the delay between two instances of fluorescence-induced energy transfer (FRET), each mediated by ribosome-tRNA interactions, was used to measure the time taken to translate codons in between the two fluorescence events (Ciryam et al. 2013). While a novel approach, the experiment measured the effect due to translating a gene from a very high-expression promoter. Additionally, the experiment was done in vitro using only a single mRNA, meaning the availability of tRNAs would be further skewed by the fact that not only is that single mRNA being translated at very high levels, but that other mRNAs which would normally be present in the cell were not available to mop up other tRNAs in the way that they would normally be demanded in a growing cell.

It is the extreme abnormality of the conditions needed to show an effect that we think may underpin a correspondence between these experimental results and our results. Let us suppose then that we can show that codon usage of a highly upregulated gene affects the translation rate. How can this observation square with the absence of higher ribosome occupancy with transcripts under normal conditions in domains rich in rare codons?

Let us consider again the balance model discussed by Qian et al. (Qian et al. 2012). They note that if highly expressed genes use codons corresponding to the most abundant tRNAs then it doesn't follow that they will be translated any faster than rare transcripts using rare codons. The key parameter is not the absolute tRNA abundance but the tRNA *availability*. If highly abundant transcripts all require the same tRNA, then this acts as a drain on the availability of that tRNA. The waiting time for a ribosome to find a rare but little in demand tRNA may then be the same (or approximately so) as the waiting time to find a "common" but much in demand tRNA. In one case the pool is small and the demand low (rare codons in lowly expressed transcripts) in the other the tRNA pool is large but the demand also large (a common codon in an abundant transcript). We have reason to expect this may be the case as tRNA abundance and codon usage have been shown to co-vary in highly expressed genes in *E. coli* (Post et al. 1979; Ikemura 1981) and yeast (Ikemura 1982). Such a correspondence has also been shown more globally within silk

worm silk gland and rabbit reticulocytes (Chavancy et al. 1979), yeast (Percudani et al. 1997) and (including at different growth rates) in *E. coli* (Dong et al. 1996). Most recently, codon usage and tRNA gene frequencies were shown to correlate across hundreds of archaeal, bacterial, and eukaryotic genomes (including human) if two basic tRNA modifications affecting binding propensities are taken into account (Novoa et al. 2012).

In fact the proposition that tRNA levels can modulate codon speeds is remarkably un-novel. Delving back in the literature, we find that in fact early theoretical work, though perhaps later ignored by many stipulating that codon speeds differ dramatically, transparently states that proportional usage between codons and tRNAs means we need not assume codons slow (Chavancy et al. 1979; Gouy and Grantham 1980; Liljenstrom et al. 1985). Other early work showed experimentally that alteration of specific tRNA concentrations can modulate chain elongation rates (Anderson 1969; Anderson and Gilbert 1969; Andersson et al. 1984). However, this fundamental observation that tRNA concentrations can modulate codon elongation rates has often been dissociated from later work which unfortunately directly transferred the findings that codons slow in the artificial settings of in vitro systems to in vivo ones. Such misplaced inference, we suggest, concomitantly bolstered the notion that the vast majority of differences in speed are due to the identity of the codons themselves, outside of any influence of tRNA concentrations.

That the presumption that certain codons slow is confused in its conception is readily apparent from the myriad of different ways in which 'slow codons' are determined: many papers cite from the (albeit problematic) experimental evidence that codons slow (discussed in Section 1), and then, satisfied the premise that certain codons slow is justified, jettison the cited experimental evidence of which particular codons may slow and move on to ad hoc definitions of which are the slowest (see e.g. Thanaraj and Argos 1996; Saunders and Deane 2010). Many such definitions revolve solely around codon usage, for example the frequency of codons (most often as they are encoded genomically) or the propensity for certain codons to be used or avoided within highly expressed genes. Although such measures correlate, it is not necessarily the case that any two metrics should incriminate exactly the same codons as being 'slow'. Alternatively, there have been efforts to use tRNA levels to define the optimality of codons (dos Reis et al. 2004; Higgs and Ran 2008; Zhang et al. 2009), however these works swing to the opposite extreme and consider tRNA concentrations while neglecting codon usage frequencies. More recently, there has been increased interest in addressing the issue of how both tRNA supply and codon demand modulate ribosome velocity (Brackley et al. 2011; Gingold et al. 2012; Qian et al. 2012), including the proposal that demand is better quantified by the transcriptomic, rather than genomic, codon usage (Pechmann and Frydman 2013).

We can then imagine an equilibrium situation in which the ribosome waiting time is the same for all codons as the demand and supply of each are balanced. Thus there is no reason to expect that rare tRNAs necessarily slow significantly (as assumed in e.g. Varenne et al. 1982), as they may be in low demand and therefore not rate-limiting. This also is consistent with our observation that, under normal growth conditions, codon usage doesn't predict ribosome occupancy. However, the same model can predict that under abnormal conditions, we might see an effect as the situation has been forced far out of equilibrium. Over-express a transcript rich in rarely used codons and the ribosome should now slow as the demand for the rare tRNAs exceeds supply. Indeed expanding the tRNA repertoire of the host genome is one strategy to maximize protein expression in an organism which has been transfected with a heterologous gene to be translated at high levels but which comprises many codons which are normally rare (correspond to rare tRNAs) in the host (Gustafsson et al. 2004). Likewise, we expect that gross modification of tRNA pools should have gross effects on translational speed as the system has been shifted away from the demand-supply equilibrium.

We find it of note that a buffering of translation speed (by a correspondence between global codon usage and tRNA pools) also seems to be recapitulated on an enzyme catalytic level when considering the translation of even a single codon. Curran and Yarus (1989) published experimental findings that the rates for charged tRNA selection at different codons span a 25-fold range. If codon usage is indeed selected for some codons to be translated quickly and others slowly, then we should expect that codons with intrinsically, mechanistically fast rates of aa-tRNA selection are enlisted for the former and codons with intrinsically slow rates of aa-tRNA selection are recruited for the latter. The authors however go on to make the following observation: that most codons whose aa-tRNAs are selected either intrinsically rapidly or slowly by the ribosome have either low or high tRNA concentrations within the cell, respectively. This would suggest that intrinsic differences in the translation speeds of certain codons are not exploited but rather normalized via their supply lines.

In evolutionary terms we expect a move out of equilibrium might occur whenever selection operates on growth rates (Ikemura 1981). Imagine a slow-growing organism comes under selection to grow faster. Under this circumstance the translational apparatus will be under selection to work faster (more ribosomes, more tRNAs). But in addition, the more abundant transcripts will come under selection to shift codon usage towards the most abundant tRNAs (whichever they may be) to free up ribosomes. These features have been observed (Rocha 2004; Higgs and Ran 2008). Once the systems returns to a supply-demand equilibrium, however, there is no reason to suppose that rare codons in lowly expressed transcripts will be processed any slower than common codons in abundant transcripts. Thus the supply-demand balance model is

consistent both with our observations and with the finding that codon bias is higher in faster growing organisms that also have more tRNAs. This means the observation that codons are used in proportion to tRNA concentrations need not be cited as evidence that codons can slow (Morlon et al. 1983; Krasheninnikov et al. 1991). Rather this observation may be indicative of the opposite, that they codons do not greatly differ in their translation speeds one to the next.

Final thoughts

The tRNA:codon balance theory described above is of course not proof in and of itself that codons do not differ *at all* in their translation speeds; rather it is a rational argument that we need not presume from the outset that codons do not do so. Given our and recent evidence that codons do not significantly differ in their elongation rates, however, we consider it likely that any speed differences between codons are of an order of magnitude that is small compared to other slowing determinants. It is almost certainly the case that codons do have different translation speeds one to the next to *some* degree, even if it be due to a stochastic search step for the correct tRNA, or differences in the binding affinities of different nucleotide pairs. The amount of selection which would have to operate on the system for every codon to have an elongation speed the same as the next would be quite extreme indeed and is difficult to imagine is actually at work within cells.

Nor can we deny the possibility that there are other complicating issues in actual translational systems leading to augmented differences in the translational speeds of certain codons, possibly for a restricted subset of codons, and under restricted conditions. The example of ribosome attenuation in bacterial amino acid operons is one such specialist case where the output of the system is clearly directed back on translation in a regulatory loop (Yanofsky 1981). For example, it has been suggested that even if tRNAs are supplied in optimal concentrations, the translation speeds of certain codons may greatly differ if their individual kinetic rate constants are sufficiently different from one another (Lizardi et al. 1979). To this we merely state that we do not find evidence, via analysis of the ribosomal footprinting dataset, that this is the case—certainly not compared to the magnitude of the positive charge effect. Another potentially complicating factor is that changes in tRNA concentrations and/or codon usage throughout the cell cycle may impact the supply:demand equilibrium. However, we note that it has been shown that a correlation between these two factors has indeed been measured at different points throughout the cell cycle in *E. coli* (Dong et al. 1996). This does not mean to say that at specific time points and precise codon locations that codon-induced slowing may not ever occur, simply that it may be unlikely that the bulk of codons at most times substantially differ in their elongation rates.

Similarly tRNAs may be differentially transcribed or modified at different cell cycle points or in different tissue types (e.g. Sharma and Borek 1970; Dittmar et al. 2006). Whether cell-cycle dependent variation in tRNA expression is a gene regulatory mechanism to increase the preferential translation of genes encoded by the corresponding codons (as suggested by Frenkel-Morgenstern et al. 2012) or a means of adapting the tRNA pool to the regulated levels of mRNAs, thereby maintaining tRNA:codon proportionality and not squandering cellular resources transcribing and processing unneeded tRNAs, is unknown. Although it has been shown that cells adjust their aminoacyl-charged tRNA acceptors during amino acid starvation to correspond to the genes most needed under these conditions (Elf et al. 2003; Dittmar et al. 2005), it is unclear whether this occurs to actively regulate gene expression via tRNA levels, or whether it is a built-in response designed to allow the cell to continue to translate protein under tRNA:codon usage equilibrium conditions without wasting energy on the creation of unneeded components of the translational machinery.

References

Ames BN, Hartman PE. 1963. The histidine operon. Cold Spring Harb. Symp. Quant. Biol. 28: 349.

Anderson WF. 1969. The effect of tRNA concentration on the rate of protein synthesis. Proc Natl Acad Sci U S A 62: 566-573.

Anderson WF, Gilbert JM. 1969. tRNA-dependent translational control of in vitro hemoglobin synthesis. Biochem Biophys Res Commun 36: 456-462.

Andersson SG, Buckingham RH, Kurland CG. 1984. Does codon composition influence ribosome function? EMBO J 3: 91-94.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. Microbiol Rev 54: 198-210.

Bonekamp F, Jensen KF. 1988. The AGG codon is translated slowly in E. coli even at very low expression levels. Nucleic Acids Res 16: 3013-3024.

Boycheva S, Chkodrov G, Ivanov I. 2003. Codon pairs in the genome of Escherichia coli. Bioinformatics 19: 987-998.

Brackley CA, Romano MC, Thiel M. 2011. The dynamics of supply and demand in mRNA translation. PLoS Comput Biol 7: e1002203.

Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. Cell 141: 355-367.

Charneski CA, Hurst LD. 2013. Positively charged residues are the primary determinants of ribosomal velocity. PloS Biol 11: e1001508.

Chavancy G, Chevallier A, Fournier A, Garel JP. 1979. Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryote cell. Biochimie 61: 71-78.

Ciryam P, Morimoto RI, Vendruscolo M, Dobson CM, O'Brien EP. 2013. In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome. Proc Natl Acad Sci U S A 110: E132-140.

Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. 2008. Virus attenuation by genome-scale changes in codon pair bias. Science 320: 1784-1787.

Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J Mol Biol 209: 65-77.

Dittmar KA, Goodenbour JM, Pan T. 2006. Tissue-specific differences in human transfer RNA expression. PLoS Genet 2: e221.

Dittmar KA, Sorensen MA, Elf J, Ehrenberg M, Pan T. 2005. Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO Rep 6: 151-157.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J Mol Biol 260: 649-663.

dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research 32: 5036-5044.

Elf J, Nilsson D, Tenson T, Ehrenberg M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. Science 300: 1718-1722.

Frenkel-Morgenstern M, Danon T, Christian T, Igarashi T, Cohen L, Hou YM, Jensen LJ. 2012. Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Mol Syst Biol 8: 572.

Gingold H, Dahan O, Pilpel Y. 2012. Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. Nucleic Acids Res 40: 10053-10063.

Goldman E, Rosenberg AH, Zubay G, Studier FW. 1995. Consecutive low-usage leucine codons block translation only when near the 5' end of a message in Escherichia coli. J Mol Biol 245: 467-473.

Gouy M, Grantham R. 1980. Polypeptide elongation and tRNA cycling in Escherichia coli: a dynamic approach. FEBS Lett 115: 151-155.

Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. Trends Biotechnol 22: 346-353.

Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol 25: 2279-2291.

Hoekema A, Kastelein RA, Vasser M, de Boer HA. 1987. Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: experimental approach to study the role of biased codon usage in gene expression. Mol Cell Biol 7: 2914-2924.

Hu X, Shi Q, Yang T, Jackowski G. 1996. Specific replacement of consecutive AGG codons results in high-level expression of human cardiac troponin T in Escherichia coli. Protein Expr Purif 7: 289-293.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146: 1-21.

Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158: 573-597.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147: 789-802.

Irwin B, Heck JD, Hatfield GW. 1995. Codon pair utilization biases influence translational elongation step times. J Biol Chem 270: 22801-22806.

Ivanov I, Alexandrova R, Dragulev B, Saraffova A, AbouHaidar MG. 1992. Effect of tandemly repeated AGG triplets on the translation of CAT-mRNA in E. coli. FEBS Lett 307: 173-176.

Komar AA, Lesnik T, Reiss C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett 462: 387-391.

Konigsberg W, Godson GN. 1983. Evidence for use of rare codons in the dnaG gene and other regulatory genes of Escherichia coli. Proc Natl Acad Sci U S A 80: 687-691.

Krasheninnikov IA, Komar AA, Adzhubei IA. 1991. Nonuniform size distribution of nascent globin peptides, evidence for pause localization sites, and a contranslational protein-folding model. J Protein Chem 10: 445-453.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science 324: 255-258.

Kurland CG. 1991. Codon bias and gene expression. FEBS Lett 285: 165-169.

Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538-541.

Liljenstrom H, von Heijne G, Blomberg C, Johansson J. 1985. The tRNA cycle and its relation to the rate of protein synthesis. Eur Biophys J 12: 115-119.

Lizardi PM, Mahdavi V, Shields D, Candelas G. 1979. Discontinuous translation of silk fibroin in a reticulocyte cell-free system and in intact silk gland cells. Proc Natl Acad Sci U S A 76: 6211-6215.

Lu J, Deutsch C. 2008. Electrostatics in the ribosomal tunnel modulate chain elongation rates. J Mol Biol 384: 73-86.

Misra R, Reeves P. 1985. Intermediates in the synthesis of TolC protein include an incomplete peptide stalled at a rare Arg codon. Eur J Biochem 152: 151-155.

Morlon J, Lloubes R, Varenne S, Chartier M, Lazdunski C. 1983. Complete nucleotide sequence of the structural gene for colicin A, a gene translated at non-uniform rate. J Mol Biol 170: 271-285.

Nakamura M, Sugiura M. 2011. Translation efficiencies of synonymous codons for arginine differ dramatically and are not correlated with codon usage in chloroplasts. Gene 472: 50-54.

Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. 2012. A role for tRNA modifications in genome structure and codon usage. Cell 149: 202-213.

Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 20: 237-243.

Pedersen S. 1984. Escherichia coli ribosomes translate in vivo with variable rate. EMBO J 3: 2895-2898.

Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol 268: 322-330.

Petersen C. 1987. The functional stability of the lacZ transcript is sensitive towards sequence alterations immediately downstream of the ribosome binding site. Mol Gen Genet 209: 179-187.

Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP. 1979. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl Acad Sci U S A 76: 1697-1701.

Qian W, Yang J-R, Pearson NM, Maclean C, Zhang J. 2012. Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. PLoS Genet 8: e1002603.

Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican A, Eaton M, Humphreys G. 1984. Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res 12: 6663-6671.

Rocha EPC. 2004. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. Genome Research 14: 2279-2286.

Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G. 1993. Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. J Bacteriol 175: 716-722.

Saunders R, Deane CM. 2010. Synonymous codon usage influences the local protein structure observed. Nucleic Acids Res 38: 6719-6728.

Sharma OK, Borek E. 1970. Inhibitor of transfer ribonucleic acid methylase in the differentiating slime mold Dictyostelium discoideum. J Bacteriol 101: 705-708.

Sørensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in Escherichia coli. J Mol Biol 207: 365-377.

Sørensen MA, Pedersen S. 1991. Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. J Mol Biol 222: 265-280.

Spanjaard RA, Chen K, Walker JR, van Duin J. 1990. Frameshift suppression at tandem AGA and AGG codons by cloned tRNA genes: assigning a codon to argU tRNA and T4 tRNA(Arg). Nucleic Acids Res 18: 5031-5036.

Stanssens P, Remaut E, Fiers W. 1986. Inefficient translation initiation causes premature transcription termination in the lacZ gene. Cell 44: 711-718.

Thanaraj TA, Argos P. 1996. Ribosome-mediated translational pause and protein domain organization. Protein Sci 5: 1594-1612.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141: 344-354.

Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12: R110.

Varenne S, Buc J, Lloubes R, Lazdunski C. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol 180: 549-576.

Varenne S, Knibiehler M, Cavard D, Morlon J, Lazdunski C. 1982. Variable rate of polypeptide chain elongation for colicins A, E2 and E3. J Mol Biol 159: 57-70.

Varenne S, Lazdunski C. 1986. Effect of distribution of unfavourable codons on the maximum rate of gene expression by an heterologous organism. J Theor Biol 120: 99-110.

Wang BQ, Lei L, Burton ZF. 1994. Importance of codon preference for production of human RAP74 and reconstitution of the RAP30/74 complex. Protein Expr Purif 5: 476-485.

Wen JD, Lancaster L, Hodges C, Zeri AC, Yoshimura SH, Noller HF, Bustamante C, Tinoco I. 2008. Following translation by single ribosomes one codon at a time. Nature 452: 598-603.

Yanofsky C. 1981. Attenuation in the control of expression of bacterial operons. Nature 289: 751-758.

Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol 16: 274-280.

IV. Positive charge loading at protein termini is due to membrane protein topology, not a translational ramp

Catherine A. Charneski & Laurence D. Hurst

Mol Biol Evol (2014) 31(1): 70-84.

Positive Charge Loading at Protein Termini Is Due to Membrane Protein Topology, Not a Translational Ramp

Catherine A. Charneski*[,1] and Laurence D. Hurst[1]

[1]Department of Biology and Biochemistry, University of Bath, Bath, Somerset, United Kingdom

***Corresponding author:** E-mail: kchartreuse@gmail.com.

Associate editor: Daniel Falush

Abstract

In the great majority of genomes, the use of positive charge increases, on average, approaching protein N-termini. Such charged residues slow ribosomes by interacting with the negatively charged exit tunnel. This has been proposed to be selectively advantageous as it provides an elongation speed ramp at translational starts. Positive charges, however, are known to orientate proteins in membranes by the positive-inside rule whereby excess charge lies on the cytoplasmic side of the membrane. Which of these two models better explains the N-terminal loading of positively charged amino acids? We find strong evidence that the tendency for average positive charge use to increase at termini is exclusively due to membrane protein topology: 1) increasing N-terminal positive charge is not found in cytosolic proteins, but in transmembrane ones with cytosolic N-termini, with signal sequences contributing additional charge; 2) positive charge density at N-termini corresponds to the length of cytoplasmically exposed transmembrane tails, its usage increasing just up until the membrane; 3) membrane-related patterns are repeated at C-termini, where no ramp is expected; and 4) N-terminal positive charge patterns are no different from those seen internally in proteins in membrane-associated domains. The overall apparent increase in positive charge across all N-termini results from membrane proteins using positive charge adjacent to the cytosolic leaflet, combined with a skewed distribution of where N-termini cross the plasma membrane; 5) while *Escherichia coli* was predicted to have a 5′ ribosomal occupancy ramp of at least 31 codons, in contrast to what is seen in yeast, we find in ribosomal footprinting data no evidence for such a ramp. In sum, we find no need to invoke a translational ramp to explain the rising positive charge densities at N-termini. The membrane orientation model makes a full account of the trend.

Key words: **translation ramp, protein topology, positive charge, N-termini.**

Introduction

Why are some amino acids, or classes of amino acid, differentially distributed within proteins? Consider, for example, the location of positively charged residues. Enrichment of positive charge nearing protein N-termini has been demonstrated in humans (Berezovsky et al. 1999), *Escherichia coli* (Berezovsky et al. 1999), and *Saccharomyces cerevisiae* (Berezovsky et al. 1999; Tuller et al. 2011). Note that while the increase in use of positive charge nearing N-termini is true on average in a given genome, it does not follow that all proteins in any given genome have positive charges in this area. A successful model should then be able to explain why some proteins do and some do not contribute to the pattern of increasing positive charge use nearing the N-terminus. Here, we consider two models that might explain this enrichment of positively charged amino acids at the starts of proteins.

The first model conjectures the positive charge enrichment is part of a ramp that controls ribosomal flow (Tuller et al. 2011). Positively charged amino acids are thought to be one (Lu et al. 2007; Tuller et al. 2011), possibly the principal (Charneski and Hurst 2013), determinant of ribosome velocity. The interior of the ribosomal exit tunnel is negatively charged (Lu et al. 2007) and positively charged residues within a protein are conjectured to interact with the negative charge in this channel, slowing ribosomal movement along transcripts (Lu and Deutsch 2008). This can explain, for example, why insertion of a long run of positively charged residues into a coding sequence stalls ribosomes (Ito-Harashima et al. 2007; Dimitrova et al. 2009). An excess of ribosomal density at the extreme 5′-ends of transcripts is present in at least one data set, which profiled the location of ribosomes along transcripts (Ingolia et al. 2009; Tuller et al. 2010). As the extent of this enrichment correlates with, among other features, the density of charged amino acids, it has been proposed that the increase in charge at the N-termini of proteins exists as one part of an adaptive speed ramp to control the flow of ribosomes at the start of translation, possibly to somehow prevent downstream traffic jams between them (Tuller et al. 2011). The notion of a ribosomal speed ramp appears to have been independently derived twice (Mitarai et al. 2008; Tuller et al. 2010), but only Tuller et al. (2011) propose a role for positive charges.

Although the translational ramp may seem an attractive explanation for N-terminal positive charge enrichment, other protein-structural origins for the use of positive charges should also be considered: just because positive charges slow ribosomes does not mean that they have been selected to do so. A more architectural hypothesis might alternately

envisage that the accumulation of positive charge at N-termini reflects some basic structural requirement of certain proteins. In this way of thinking, positive charge is not selected for because of its influence on a short-lived process (translation), but because of its contribution to the integral composition or structure of the protein itself. As positive charges have been well established to play a role in determining the orientation of integral membrane proteins, we here consider their usage as a possible alternative explanation for the N-terminal enrichment of positive charge.

The so-called positive-inside rule, which applies to proteins in both prokaryotes and eukaryotes, both with and without signal sequences, says that proteins orientate so that excess positive charge near hydrophobic membrane-spanning regions lies on the cytoplasmic side of the membrane (von Heijne and Gavel 1988; Sipos and von Heijne 1993). Correspondingly, the experimental addition of positively charged residues to normally periplasmic regions is capable of inverting the topology of a protein, such that the excess of positive charges will lie in the cytosol (Nilsson and von Heijne 1990). The insertion of proteins into membranes is thought to be achieved by a variety of conserved translocases and integrases (such as the well-described Sec translocon) acting both independently and cooperatively (Samuelson et al. 2000; Dalbey et al. 2011; Nishiyama et al. 2012). The addition of positive charges to the N-termini of transmembrane proteins can prevent the translocation of the termini across membranes in both *E. coli* and eukaryotes (Gafvelin et al. 1997), whether they require the main Sec protein-conducting channel (Li et al. 1988; Yamane and Mizushima 1988) or not (Whitley et al. 1994).

Although the prevalence of the positive-inside rule is recognized, the mechanisms by which positive charges exert their topogenic effects are not well understood. Membrane protein topology may arise, at least in part, from positive charges near hydrophobic stretches stopping the transfer of further stretches of the protein across the membrane, and thus anchoring the hydrophobic region within the bilayer (Kuroiwa et al. 1990). The positively charged residues might electrostatically interact with the negative phospholipid groups of the bilayer, preventing translocation of this portion of the protein through the membrane (Gallusser and Kuhn 1990; van Klompenburg et al. 1997). The proton-motive force leading to the acidification of the periplasm may be required for the translocation of some protein segments, facilitating transfer of negative but not positive residues across the membrane (Whitley et al. 1994; Kiefer et al. 1997; Delgado-Partin and Dalbey 1998). The arrangement of conserved positive and negative charges within the exoplasmic and cytoplasmic portions, respectively, of the Sec translocon itself could additionally contribute to the topogenesis of membrane proteins by interacting with charged residues within the proteins (Goder et al. 2004).

Can the positive-inside rule alone explain the differential location of positively charged amino acids within proteins or do we in addition need to evoke selection on ribosomal velocity? The positive-inside rule makes numerous predictions regarding which proteins and where in the proteins we expect to see positive charge enrichment. We test these predictions and show that the increase in average positive charge usage at the start of proteins is parsimoniously explained in full as a consequence of the need for many proteins to thread themselves through and orientate themselves in lipid bilayers. In both *E. coli* and *S. cerevisiae*, we find increasing N-terminal charge among membrane proteins, not cytoplasmic ones. Focusing on *E. coli* (due to the need for large sample sizes of experimentally supported transmembrane protein annotations), we find positive charge enriched at the point where cytosolically exposed N-termini enter the membrane, in accordance with the positive-inside rule. That similar patterns are repeated at the C-terminus, where no ramping effect on downstream translation is to be expected, suggests the N-terminal positive charge pattern is purely protein-structural in origin. Cleavable signal sequences in *E. coli* tend to be rich in cations (von Heijne 1984) and lend an additional enrichment of positive charge at protein starts. We finally demonstrate that N-terminal positive charge patterns can be entirely explained by patterns of downstream cation usage in proximity to membranes. Thus, the overall increase in positive charge across all N-termini results from the use of positive charge adjacent to the cytosolic leaflet of membranes combined with a skewed distribution of where cytosolic N-termini cross the plasma membrane.

Results

Across All Three Domains of Life, Average Positive Charge Usage Increases Nearing Protein N-Termini

First, we ask about the generality of the N-terminal loading of positively charged amino acids. Although an increase, on the average, at N-termini of the density of positively charged amino acids has been seen in a few species, just how general is it? Upon aligning proteins by their N-terminus and calculating the average usage of positive charge within a given amino acid site, we observe that the average use of positive charge within 622 of 648 organisms (including the vast majority of bacteria and archaea studied) tends to increase nearing the N-terminus (fig. 1; supplementary fig. S1, Supplementary Material online). Given our constraints as regards which coding sequences we will include (see Materials and Methods, Sequences), we were only able to retain a small number of proteins for analysis for some eukaryotes, approximately 1–10% of the total number of genes encoded in the genome (supplementary fig. S1, Supplementary Material online). We consider it a strong possibility that the positive charge pattern is not seen in these organisms (supplementary fig. S1, Supplementary Material online) due to the sequencing quality of these genomes. Indeed, it is only for such low-coverage eukaryotes that we do not observe significantly enriched N-terminal charge, quite possibly because we had to remove (via our sequence filters, see Materials and Methods) the subset of proteins that contribute to this pattern when all proteins are considered within an organism en masse.

This increasing use of positive charge near N-termini in 622 species is consistent with prior observations that mean charge

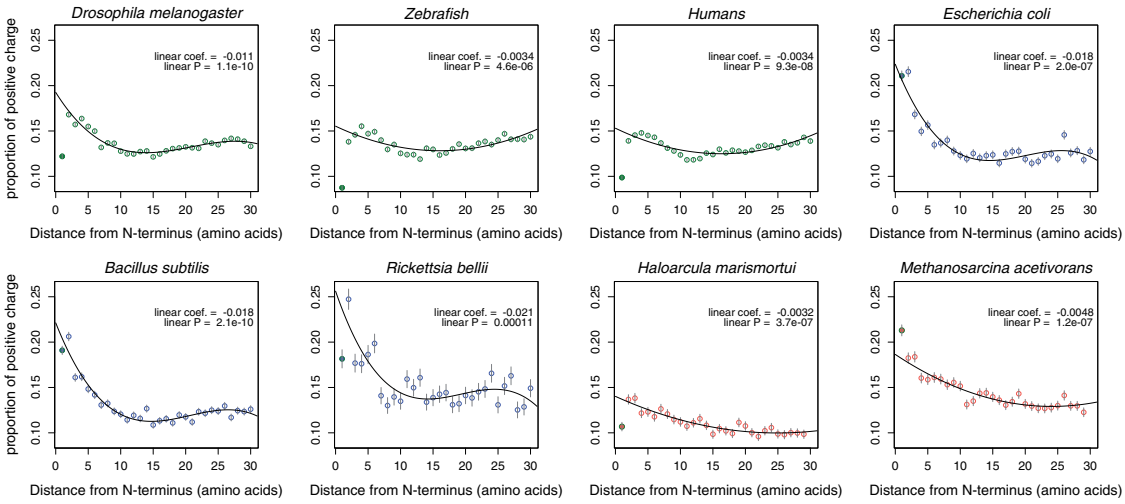

Fig. 1. Average positive charge usage in an organism increases toward the N-terminus across all three domains of life. Whether the use of positive charge increases nearing the N-terminus is determined by the sign of the linear coefficient term (see Materials and Methods). Representative genomes are shown from eukaryotes (*Drosophila,* zebrafish, and humans), bacteria (*E. coli*, *B. subtilis, and R. bellii*), and archaea (*H. marismortui*, *M. acetivorans*). See supplementary figure S1 (Supplementary Material online) for more plots from 648 species. Here and in all plots, error bars represent the standard error of the mean. The first amino acid following the initiating methionine (filled points) is often an outlier and was ignored in all regressions.

increases nearing the N-terminus in *S. cerevisiae* and *E. coli* (Tuller et al. 2011). As we are interested in the potential ramifications of positive charge on ribosomal slowing, however, and we previously found no effect of negative charge on translation speed (Charneski and Hurst 2013), we consider only positive, not negative, charge here and in all further analyses.

Increasing N-Terminal Average Positive Charge Is Not Found in Nonmembrane Proteins

Investigating the positive charge pattern among groups of proteins that are differentially sublocalized within *E. coli* and *S. cerevisiae* shows that positive charge does not generally increase nearing N-termini in cytosolic proteins but in proteins that are localized near to or potentially within membranes. In *E. coli*, increasing N-terminal charge is found among proteins generally localizing to the inner and outer membranes as well as periplasm, and in yeast, such a pattern is found in proteins resident near the mitochondrion, endoplasmic reticulum (ER), Golgi, and vacuole (fig. 2; see supplementary fig. S2 [Supplementary Material online] for more yeast subcellular localizations). Hence, the proteomic-scale pattern in *E. coli* and *S. cerevisiae* in figure 1 results from the locations of positive charges in a subset of proteins within the organisms.

The Increased Positive Charge at N-Termini Is Associated with Both the Topology and Sometimes Signal Sequences of Transmembrane Proteins

What is it about membrane proteins that lead to increasing N-terminal positive charge (fig. 2)? As the orientation of membrane proteins correlates with the density of positive charges on the cytoplasmic side of the membrane (von Heijne and Gavel 1988; Sipos and von Heijne 1993), we wondered whether the rise in positive charge at N-termini may be linked to the orientation of these termini. Indeed, we find that among transmembrane proteins in both *E. coli* and *S. cerevisiae*, cytoplasmically orientated N-termini show a far greater increase in positive charge at the tail than do those in the periplasm (fig. 3). We also note that cleavable signal sequences in *E. coli* tend to be positively charged (fig. 3*A*), as previously reported (von Heijne 1984). This is in line with findings that such charges in cleavable signal sequences, while not always essential for export, can significantly enhance the rate of translocation (Vlasuk et al. 1983; Puziss et al. 1989). This means that periplasmic proteins in *E. coli* display even more minimal N-terminal charge when proteins with N-terminal signal sequences are excluded (fig. 3*A*, last panel).

To determine whether the enrichment of positive charge in transmembrane protein N-termini that are cytosolic is significantly different from those that are periplasmic, outside of any additional contribution of signal sequences, the following randomization was performed independently in *E. coli* and yeast. Signal-less proteins with N-termini in the cytosol and periplasm were combined into one group and then randomly sampled without replacement into two groups the same sizes as the observed groups. For each resampling, the average proportion of positive charge at each position in the first 30 amino acids was calculated in each of the two resampled groups. We then summed the differences in the average proportion of positive charge usage between the two sets at each N-terminal amino acid position using a one-tailed approach, as we have a strong prior that the cytoplasmic termini will display greater positive charge. If linear fits for the randomized N-cytosolic and N-periplasmic intersected before 30 amino acids downstream of the N-terminus, we stopped summing

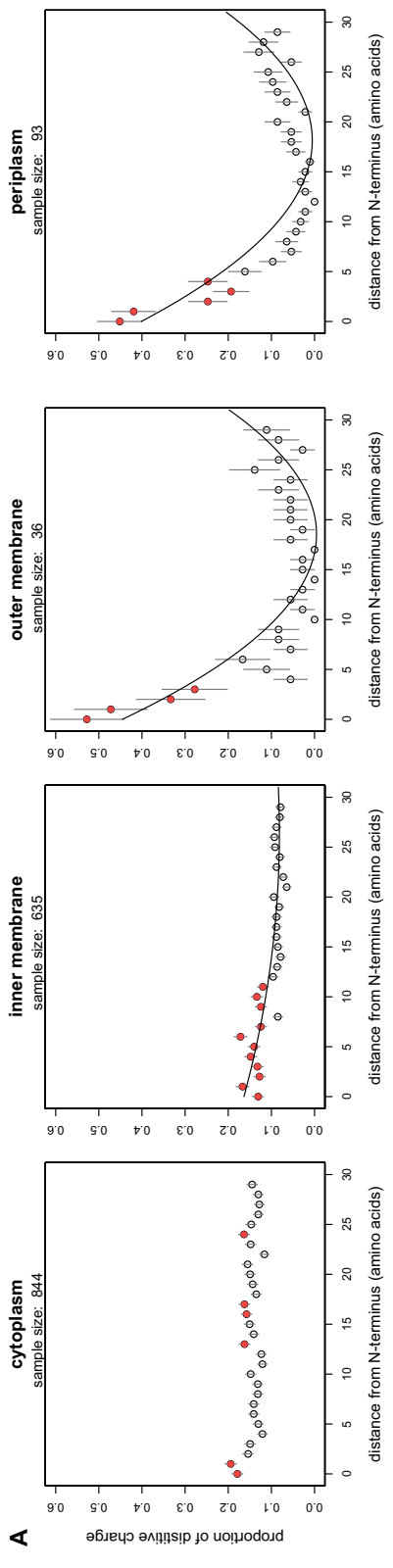

FIG. 2. Only proteins potentially associated with membranes, not cytosolic proteins, show increasing N-terminal charge. Note that locations in this figure correspond to general subcellular locations of entire proteins, but say nothing specifically about the exact locations of protein termini, which in the case of transmembrane proteins may vary depending on their orientation in the membrane. In addition, proteins classified as "membrane" proteins may for example be only peripherally bound (Han et al. 2011). Filled points show positions of significantly enriched ($P < 0.05$) positive charge (see Materials and Methods). Linear versus quadratic best fits were determined by ANOVA of nested models. Whether y is increasing approaching the end of the protein ($x = 0$), in the quadratic regressions, is determined by the sign of the x-term coefficient (see Materials and Methods). (*A*) *Escherichia coli*: Cytosol: no increase in charge at N-termini: regression of $y \sim x$, slope $P = 0.20$. Outer membrane and periplasm: regression of $y \sim x^2 + x$, fitted x-term value of -0.049 ($P = 5.4e{-}09$), $r^2 = 0.74$; -0.044 ($P = 1.15e{-}10$), $r^2 = 0.80$, respectively. Inner membrane: regression of $y \sim x^2 + x$, fitted x-term value of -0.00622 ($P = 0.00017$), $r^2 = 0.68$. (*B*) *Saccharomyces cerevisiae*: For all regressions, the point at $x = 0$ was excluded as it is often an outlier (see also supplementary fig. S2 [Supplementary Material online] for more subcellular locations in yeast). Cytosol: no increase in charge at N-termini: regression of $y \sim x$, slope $P = 0.66$. ER and vacuole, regression of $y \sim x^2 + x$, fitted x-term value of -0.0043 ($P = 0.033$), $r^2 = 0.10$; $-1.1e{-}02$ ($P = 0.0012$), $r^2 = 0.32$, respectively. Golgi, regression of $y \sim x$, linear coefficient -0.0032, $P = 0.010$, $r^2 = 0.19$.

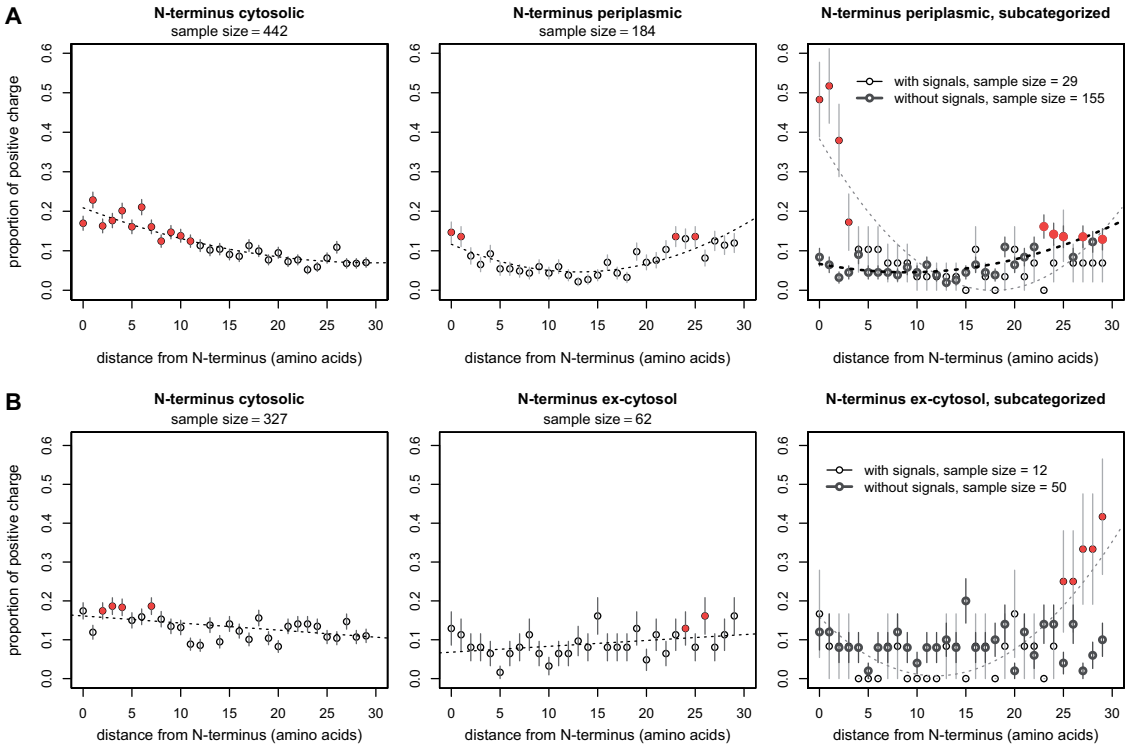

Fig. 3. The topology and signal sequences of transmembrane proteins cause N-terminal positive charge loading. All proteins considered in these plots are transmembrane. Linear versus quadratic fits were determined by ANOVA of nested models. Whether y is increasing approaching the end of the protein ($x = 0$) any order of regression is determined by the sign of the linear term coefficient (see Materials and Methods). Filled points show positions of significantly enriched ($P < 0.05$) positive charge (see Materials and Methods). Rows *A* and *B* have different y axes. (*A*) *Escherichia coli*: All regressions are of the form $y \sim x^2 + x$. N-termini in the cytosol: with signal sequences, x-term value of −0.0093 ($P = 5.0e{-}06$), $r^2 = 0.82$. There are no proteins with cytosolic N-termini which have signal sequences. N-termini in the periplasm: all proteins, fitted x-term value of −0.011 ($P = 6.6e{-}06$), $r^2 = 0.61$; only those with signal sequences, x-term −0.043 ($P = 6.8e{-}08$), $r^2 = 0.69$; those without signal sequences, x-term coefficient −0.0048 ($P = 2.5e{-}02$), $r^2 = 0.62$. (*B*) *S. cerevisiae*: N-termini in the cytosol: $y \sim x$ slope −0.0018 ($P = 0.0029$), $r^2 = 0.25$. N-termini ex-cytosol: $y \sim x$ slope 0.0015 ($P = 0.045$), $r^2 = 0.10$; only those with signal sequences, $y \sim x^2 + x$ regression x-term −0.025 ($P = 2.6e{-}05$), $r^2 = 0.74$; those without signal sequences, $y \sim x$ slope $P = 0.94$.

the differences at the point where the fits intersected; otherwise, we summed the differences over all 30 N-terminal amino acid positions. After 10,000 iterations, P was calculated as $(m + 1)/(n + 1)$, where n is the number of iterations and m is the number of times the randomized "area between the curves" was greater than or equal to that observed. This test indicates the chance of randomly obtaining such a large difference between the two curves in similar sized groups given the transmembrane proteins used to calculate those curves is rather low indeed in both *E. coli* ($P = 0.0001$) and yeast ($P = 0.0001$).

Thus, we conclude the difference in positive charge usage between the two groups is highly significantly different, with positive charge loading occurring in the cytosolic N-termini of integral membrane proteins (in agreement with the positive-inside rule). This observation is straightforwardly interpreted in terms of the protein biochemistry/membrane orientation argument. In principle, if one could propose a post hoc rationalization as to why membrane proteins in particular uniquely require the putative ribosomal slowing effects of positively charged residues, then this result can also be considered as not falsifying the positive charge ramp model. We are unaware, however, of any such post hoc rationalization.

Positive Charge Is Enriched in Cytosolic N-Tails near and Just up to the Point Where the Proteins Enter the Membrane

In the previous section, we show that increased N-terminal positive charge is associated with an N-cytosolic transmembrane topology. We now look more closely at the configuration of cytoplasmic N-tails in relation to the plasma membrane and investigate where positive charge tends to be used in relation to the point where these tails penetrate the membrane.

Although it would require a more specialized hypothesis to imagine a scenario in which only N-cytosolic membrane proteins require a positive charge driven ramp, we assume for the moment that such a hypothesis is possible for the sake of the argument. If positive charge is enriched in cytosolic N-termini (within the first 30 amino acids) to slow ribosomes, we might expect that within a given protein, positive charge tends to be

used closer to N-termini than to the downstream region where the protein enters the membrane—or, perhaps we might expect no correlation at all between the densest regions of positive charge and their proximity to the membrane. However, if N-termini are enriched in positive charges to orientate proteins in membranes, we expect to see a bias in positive charge usage close to the point where the protein enters the membrane (von Heijne and Gavel 1988), with less positive charge usage upstream, closer to the initiating methionine. We examined the N-cytosolic regions of transmembrane proteins that were 10–30 amino acids in length and compared the density of positive charges in the first five amino acids at the N-terminus (following the first amino acid, normally an uncharged methionine) with the density of positive charges in the five downstream cytosolic amino acids adjacent to the membrane. (The ten amino acid minimum for these protein tails simply gives enough length to allow us to distinguish upstream or N-terminal amino acids from downstream, membrane-adjacent ones.) Not only do we find that 81% of proteins investigated have more positive charge in their membrane-adjacent region than upstream at the N-terminus (binomial test, $P < 2.2e{-}16$) but also we find that the magnitude of positive charge in this membrane-adjacent region is significantly higher than in the upstream N-terminal region within the same protein (paired one-sided Wilcoxon test, $P < 2.2e{-}16$). Thus, positively charged residues are used in proximity to the plasma membrane and not in proximity to the N-terminus per se.

We also find that the last positive charge used in an N-cytosolic segment tends to lie just near the face of the plasma membrane (fig. 4). We however find no such trend for positive charge usage for N-termini that lie on the periplasmic side of the bilayer (fig. 4). These findings are consistent with the above proposition that positive charge use at N-termini is linked to membrane proximity and the positive-inside rule (Heijne 1986).

The Degree of Positive Charge at the N-Terminus Corresponds to the Length of Transmembrane Peptide Exposed to Cytosol

That the average increase in positive charge at cytosolic N-termini is actually a function of the point where individual proteins intersect the membrane and is not a feature inherent

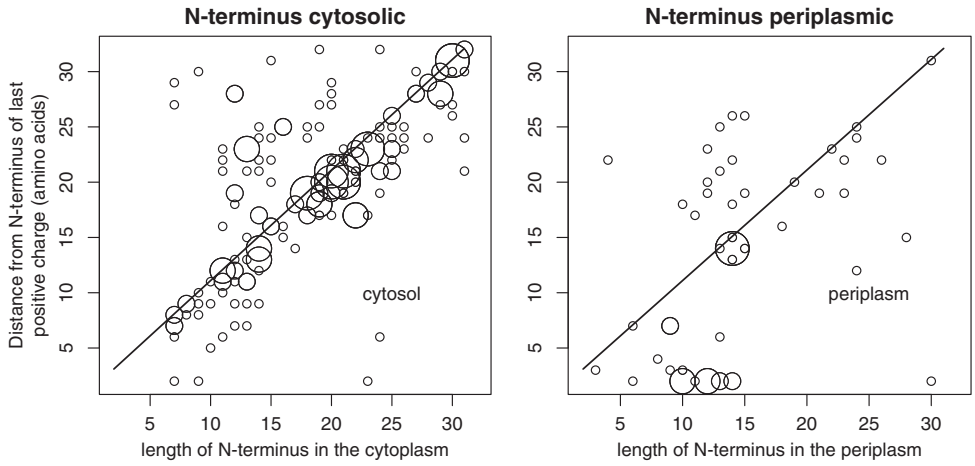

Fig. 4. Among *Escherichia coli* transmembrane proteins, the last positively charged amino acid of cytoplasmic N-termini tends to lie near the inner leaflet of the membrane. The size of the point is proportional to the number of times that point is plotted. The length of the N-terminal fragment must be less than or equal to 30 residues, purely because this is the length of the major phenomenon we are trying to investigate (see fig. 3A, N-cytosolic proteins). In either plot, if the use of positive charge is closely associated with membranes we should expect dense points near the line $x = y$ which represents the *face* of the appropriate membrane (the membrane itself will occupy more space above the thin line depicted). We note that the TOPDB protein topologies used in making this figure are supported by experimental evidence and hence the trends we report here are not an artifact of prediction algorithms (see Materials and Methods). N-terminus cytosolic: The point of the next membrane crossing—that is, where the N-terminus exits the membrane into the periplasm—must occur at least 31 residues downstream of the start of the protein, so as to not interfere with the N-terminus-into-cytosolic leaflet transition we wish to inspect. The diagonal line represents the inner face of the inner membrane and is depicted for visual purposes only. Spearman's rho between x and y, 0.67, $P < 2.2e{-}16$; the slope of a standardized major axis regression of $y \sim x$ is not significantly different from 1 ($P = 0.18$; slope coefficient 95% CI: 0.97, 1.2). Binomial test that positive charges have a 50/50 chance of being found on either side of the inner leaflet of the inner membrane, $P < 2.2e{-}16$ (with 156 out of 194 observations located leading up to and just at the cytosolic side of the membrane). N-terminus periplasmic: Proteins with signal sequences are excluded from the plot as we wish to investigate the remaining interaction of the protein with the membrane once they are cleaved. Similar to above, the point where the N-terminus exits, the membrane into the cytoplasm must occur at least 31 residues downstream of the start of the protein. The diagonal line represents the outer face of the inner membrane and is depicted for visual purposes only. Spearman's rho between x and y, 0.47, $P = < 0.00033$; the slope of a standardized major axis regression of $y \sim x$ is significantly different from 1 ($P = 0.0057$; slope coefficient 95% CI: 1.1, 1.8). Binomial test that positive charges have a 50/50 chance of being found on either side of the inner leaflet of the inner membrane, $P = 0.00018$ (with 41 out of 54 observations on the periplasmic side of the membrane).

to the N-termini specifically is well demonstrated visually. Upon progressively increasing the maximum length of N-cytosolic tails to be plotted, we see that the area of the N-terminus over which average positive charge increases is a function of the length of the exposed cytosolic tail (fig. 5*A*). This is in line with our finding that positive charges are used in the cytosolic portion of the protein before contacting the membrane. However, when we consider independent ranges of N-cytosolic lengths, it becomes apparent that positive charge is in fact not used more heavily in all proteins near the very beginning of proteins but at the point where the protein meets the membrane (fig. 5*B*). Thus, the positive charge curve for all N-cytosolic proteins (fig. 3) appears to increase because the distribution of tail lengths is weighted toward the shorter end, with the majority of N-cytosolic tails being quite small (supplementary fig. S3, Supplementary Material online). When combined with a tendency for higher positive charge usage in proximity to membranes, this length distribution creates the monotonic curve seen in figure 3 (see fig. 6 for a graphical representation of this concept). This finding strongly argues for the protein orientation argument and against the ramp argument, as the ramp would propose (we presume) that all proteins should have the positive charges either randomly scattered or in approximately the same place.

Positive Charge Usage Is Also Tied to Transmembrane Protein Architecture at the C-Terminus

We consider that if similar trends in positive charge use exist at C-termini, where no ramping effect on downstream translation should be expected, this would be strong evidence that N-terminal positive charge usage is a consequence of protein biochemistry rather than translational regulation. Indeed, we find that increasing positive charge usage nearing membrane protein C-termini is strong among those that lie in the cytosol and remarkably minimal in those that are periplasmic (supplementary fig. S4, Supplementary Material online). Among the transmembrane proteins that have between 10 and 30 amino acids in the cytosol at the C-terminus, more positive charge is found within the five amino acids closest to the membrane on the cytoplasmic side compared with the five most C-terminal amino acids (binomial test, $P = 0.016$), ignoring the last two amino acids of proteins because their basicity can greatly enhance translation termination efficiency (Mottagui-Tabar et al. 1994; Bjornsson et al. 1996). Additionally, this density of positive charge in the five amino acids just adjacent to the cytoplasmic face of the membrane is significantly greater than the magnitude of positive charge in the corresponding C-terminal region (paired one-sided Wilcoxon test, $P = 3.7e{-}05$). As might be expected, if increasing positive charge usage at C-termini is tied to orientating proteins within membranes, the most upstream positively charged residue within the last 30 amino acids of a protein lies very close to the inner leaflet of the membrane (supplementary fig. S5, Supplementary Material online). Similar to the N-terminus, the degree of positive charge at the C-terminus is a function of the length of transmembrane tail that is exposed to the cytosol, with a combination of the lengths of C-tails exposed to the cytosol (supplementary fig. S3, Supplementary Material online) and a tendency for positive charge to be used near membranes contributing to the emergent increasing charge pattern seen in C-cytosolic membrane protein termini (supplementary fig. S6, Supplementary Material online).

Given these results, we conclude that transmembrane protein topology is capable of creating increasing average positive charge curves at either terminus, simply as a consequence of the membrane protein topologies at that terminus, without the need to invoke an N-terminal translational speed ramp to explain positive charge use at the beginnings of proteins.

Positive Charge Usage in Proximity of Transmembrane Regions Can Entirely Account for the Increasing Positive Charge at the N-Terminus

That similar phenomena contribute to increasing average positive charge nearing both N- and C-termini strongly suggests that N-terminal average positive charge patterns are caused by selection on transmembrane protein structures alone. As an additional control that no other major force is contributing to the increase in positive charge at N-termini, we asked whether N-terminal positive charge usage near membranes substantially differs from that observed in proximity to membrane crossings that occur in the middle of the protein. Such transmembrane regions that lie further downstream in the protein sequence allow us to measure membrane-proximal positive charge usage patterns outside any extra influence on positive charge usage at the N-terminus.

We located all N-cytosolic into membrane transitions that occurred more than 45 amino acids downstream of the protein start. This allowed us to create a profile of positive charge usage near these downstream membrane crossings. For each detected transition, we located six adjacent windows of five amino acids each: the first three windows lying in the cytoplasm and the latter three windows in the membrane. We calculated the number of positive charges present in each window in each eligible protein, allowing us to eventually calculate the average density of positive charge in each downstream window position relative to the membrane. These average densities were then used to "reconstruct" the upstream (first 30 amino acids) N-terminal positive charge. For every protein that went into making figure 3 (N-cytosolic proteins panel), the point at which the N-terminus hits the membrane was recorded, and the reconstructed positive charge for that protein was incremented in each possible 5-amino acid window surrounding that point by the observed average density in that window position relative to the membrane. The observed N-terminal average positive charge pattern (fig. 3, N-terminus cytosolic) is not significantly different from this reconstructed N-terminal positive charge resulting from patterns of downstream positive charge usage patterns combined with the locations of where N-termini cross from the cytoplasm into membranes (fig. 7). Thus, we infer that membrane protein topology alone is responsible for the increase in positive charge at the N-terminus. Our ability to reconstruct the increasing positive charge pattern in the

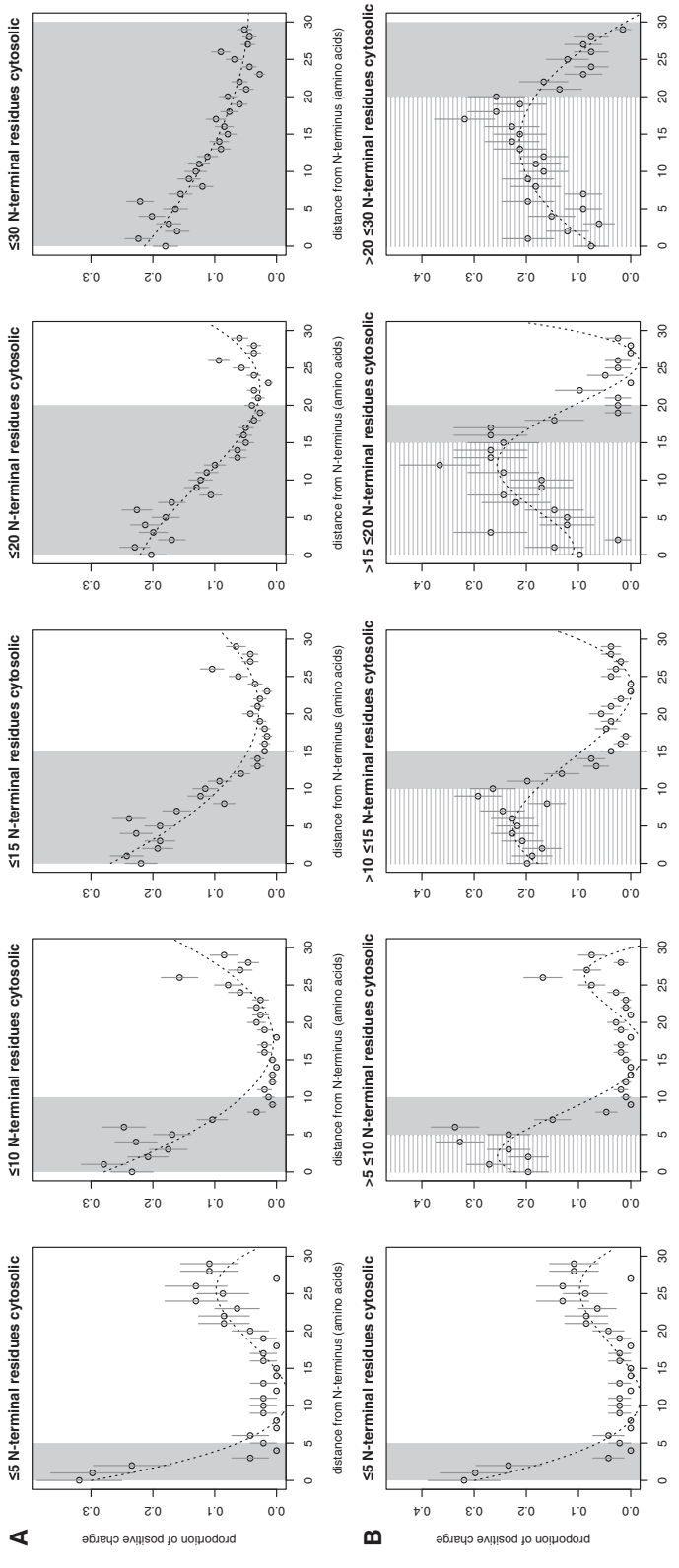

FIG. 5. The degree of positive charge at the N-terminus corresponds to the length of transmembrane peptide exposed to the cytosol. All data are from *Escherichia coli*, and all proteins considered in plots are transmembrane. Higher-order best fits were determined by ANOVA of nested models. The first panels in rows *A* and *B* are the same. (*A*) Proteins with a maximal length of the N-terminus within the cytosol are considered; hence, all subsequent plots encompass data from the previous plots within this row. The solid shaded region on each plot represents the possible range of locations wherein all the N-cytosolic proteins considered in that plot must cross into the plasma membrane. In other words, all the cytosolic portions in these plots reside within the solid-shaded regions, but the shaded regions may also contain noncytosolic residues if the protein crosses into the inner membrane before the shaded region ends. Average positive charge usage is well predicted by the propensity for the N-terminus to be cytosolic, as shown by a regression of point of inflection on *x* axis where positive charge usage is lowest ~ no. cytosolic N-terminal residues: slope = 0.79, slope $P = 0.002$, $r^2 = 0.96$. (*B*) Each plot considers a range of residue lengths that are exposed to the cytoplasm at the N-terminus, that is, the plots in this row consider mutually exclusive sets of proteins. The solid shaded region, as in row *A*, represents the range of locations where the N-cytosolic protein transitions into a membrane. The region with striped, lighter shading represents the range of residues which can be guaranteed to reside within the cytosol. When considered this way, it becomes apparent that positive charge is used, on average, more and more just up to the point where the protein meets the inner face of the (negatively charged) phospholipid bilayer: regression of point of maximal positive charge usage along *x* axis ~ minimum no. of cytosolic N-terminal residues: slope = 0.75, slope $P = 0.006$, $r^2 = 0.93$.

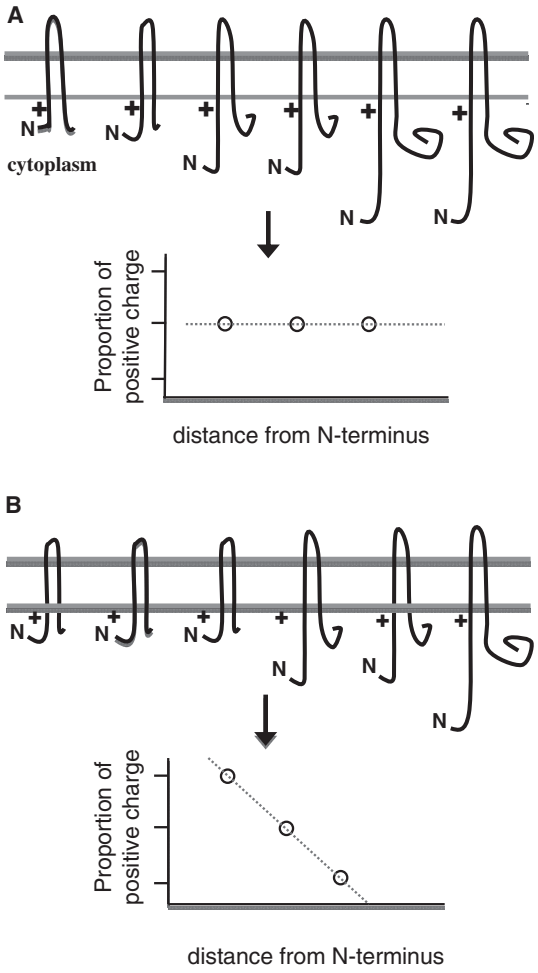

Fig. 6. Illustration of how bias in where N-termini cross membranes coupled with positive charge use near the cytosolic leaflet can cause increasing positive charge use nearing the N-terminus. Only a single positive charge is shown on each protein for sake of diagrammatic clarity. (*A*) Each of three N-terminal lengths is equally represented, leading to a slope of zero on the resulting regression line. (*B*) The locations of where N-termini exit the cytosol are skewed such that the shorter N-terminal length is overrepresented and the longer N-terminal length is underrepresented. This causes an apparent increase in positive charge use at the N-terminus within the "average protein."

"average protein" is an additional demonstration that such a pattern can and does indeed result from positive charge in transmembrane regions adjacent to the cytosolic leaflet in conjunction with a bias for short N-terminal lengths (fig. 6).

Do Membrane Proteins Have More Ribosomes on the N-Termini?

The hypothesis that we were testing provided evidence that positive charges at N-termini correlated with increased ribosomal loading at transcript starts in yeast (Tuller et al. 2011). Do we then find increased ribosomal occupancy in membrane proteins as our findings might predict? Examining the relative change in ribosomal occupancy along a transcript (relative to the average occupancy of that transcript; see Materials and Methods), we find that ribosomal density is enriched, particularly over the first several codons in transmembrane proteins compared with cytosolic ones (fig. 8). Another study similarly found an enrichment of ribosomal footprints in ER-associated proteins compared with cytosolic ones in a human embryonic kidney cell line (Reid and Nicchitta 2012). However, we observe that among membrane proteins, N-periplasmic proteins are either at least as or more enriched in ribosomal density along the first few codons of a transcript than N-cytosolic ones in both *E. coli* and *S. cerevisiae* (fig. 8), suggesting that increasing average positive charge density is not responsible for the most proximal ribosomal densities (over the first few codons) observed on transcripts. After this initial excess, membrane proteins with cytosolic N-termini do appear to stay somewhat more occluded by ribosomes than other proteins, consistent with the result that positive charges correlate with average ribosomal density in yeast (Tuller et al. 2011). However, ramp-like densities are observed in all subclasses of protein (fig. 8), indicating some feature common to all subclasses is responsible for the bulk of the 5′ ribosomal loading.

We have thus far assumed that the ramp observed in yeast is, like the increased positive charge at N-termini, possibly a phylogenetic universal. However, no such ramp was observed in mouse embryonic stem cells (Ingolia et al. 2011). What then about *E. coli*? Given increased usage of three features associated with ribosomal slowing (charge, codon bias, and RNA folding) in the first 30 + codons of *E. coli* proteins, much as seen in yeast, it was presumed that *E. coli* would also have a ramp-like slowing effect (Tuller et al. 2011, their figure 1). However, scrutiny of figure 8 indicates that, in contrast to yeast, apart from the ribosomal excess over the initial (first few) codons, there is no evidence for an extended ramp in *E. coli*, either in membrane proteins or cytoplasmic ones (fig. 8). Indeed after the initial ribosomal excess ($x > 4$) in *E. coli*, occupancy at each position is roughly the same as the average occupancy along the rest of the gene, and actually tends to increase somewhat from $x = 5$ to $x = 30$, counter to the above prediction (see legend of fig. 8).

Discussion

We find that the increasing use of positive charge nearing protein N-termini, seen when averaging over all proteins in a proteome, is due to transmembrane protein topology in both *E. coli* and *S. cerevisiae* (see fig. 6 for illustration). Such a finding is in accordance with positive charges being used to orientate N-tails in the cytosol as opposed to periplasm. The hypothesis that positive charge use at N-termini is due to membrane protein orientation makes correct predictions about which proteins have positively charged N-termini and where in proteins enrichment of positive charge is seen. Although, on average, positive charge use increases approaching N-termini (figs. 1–3), in fact positive charge is used closer to membrane intersection point than to the N-terminus proper (figs. 4 and 5). Hence, the overall increasing average positive

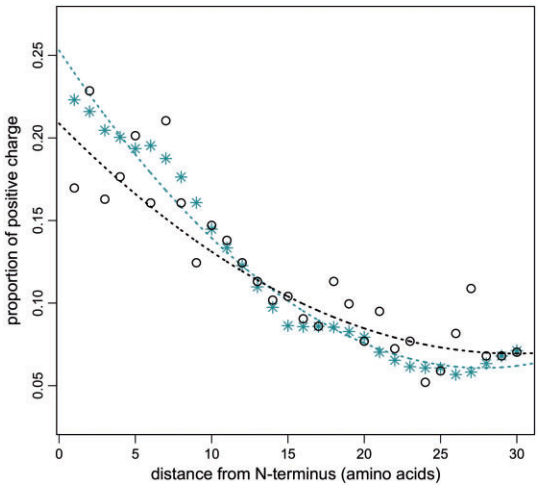

Fig. 7. The N-terminal positive charge pattern in *Escherichia coli* can be entirely explained by patterns of positive charge usage near transmembrane regions. Asterisks: N-terminal positive charge reconstructed according to patterns of positive charge usage seen in the vicinity of transmembrane segments which are further downstream (see Results). Regression of $y \sim x^2 + x$, x-term coefficient −0.014, $P = 5.6e{-}16$. Circles: observed positive charge in cytosolic N-termini (as in fig. 3*A*, first panel) plotted for comparison. Regression of $y \sim x^2 + x$, x-term coefficient −0.0093, $P = 5.0e{-}06$, $r^2 = 0.82$. The observed and reconstructed positive charge usage are not significantly different (paired t-test: $P = 0.78$; the differences between paired observed and reconstructed charge are normally distributed: Shapiro test, $P = 0.083$).

charge pattern at protein starts is created by the length distribution of N-cytosolic tails of transmembrane proteins (fig. 5). That similar phenomena contribute to increasing average positive charge nearing C-termini strongly suggests that N-terminal charge patterns are simply a consequence of the structural needs of proteins, namely to orientate themselves in membranes in accordance with the positive-inside rule. That N-terminal average positive charge patterns can be entirely reconstituted from downstream positive charge usage patterns near membranes (fig. 7), where no selection on either ramping or other potential reasons for charge selection which might be particular to the N-terminus, further confirms the protein-structural basis for this positive charge pattern. Importantly, we find no need to invoke a translational ramp to explain N-terminal positive charge densities.

Our results do not preclude that positive charges may be selected at termini for other physiochemical reasons. For example, cytoplasmically located proteins, while not displaying increasing charge nearing N-termini, do not show an absence of positive charge either (fig. 2). It is possible that within a subset of proteins (either transmembrane or cytosolic) an exposed tail may need positive charges to, among other things, bind other groups, including negative charges in nucleic acids (Moarefi et al. 2000), or that positive charge may be selected for the exposed termini residues to enhance protein solubility (Islam et al. 2012).

Our results, moreover, should not be overinterpreted. Our analysis is not designed to ask whether the ramp (as observed in yeast) is real or whether the ramp is adaptive. We simply wish to know whether the increase in average positive charge nearing N-termini, which we have shown to be a widespread, if not phylogenetically universal pattern, is best explained as part of a mechanism to stall ribosomes. We cannot on the basis of our results conclude that there is no ramp in any organism or that any potential ramp is necessarily not adaptive. We note, rather, that the loading of positive charges at N-termini is not evidence that such stalling is adaptive or that positive charges are selected at N-termini for gene regulatory purposes, especially as we find a more parsimonious explanation for the presence of these charges. Consistent with positive charge being better explained by a factor other than a regulatory ramp, we find no evidence for a 5′ ribosomal ramp of the predicted dimensions (at least 31 codons) in any class of protein upon examination of ribosomal footprinting data in *E. coli* (fig. 8). This is a surprising result as 5′ aberrations in codon usage (Eyre-Walker and Bulmer 1993), encoded positive charge (Berezovsky et al. 1999), mRNA folding (McCarthy and Bokelmann 1988; de Smit and van Duin 1990), or a combination of the three were predicted to cause a 5′ increase ribosomal densities along a transcript (Tuller et al. 2011).

Taken together with other recent results, however, our findings add a little to the literature questioning the validity of the adaptive ramp hypothesis. Some have questioned, as we did, whether features of the ramp are better explained in other terms. The hypothesized ramp is posited to be, in addition to a consequence of charge, caused by two other properties of 5′-ends of mRNAs: nonoptimal codon usage and strong RNA folding (Mitarai et al. 2008; Tuller et al. 2010). Leaving aside the problem that analysis of ribosome protection data failed to find evidence that nonoptimal codons slow ribosomes under normal conditions (Qian et al. 2012; Charneski and Hurst 2013), there is an alternative and more parsimonious interpretation of the enrichment of rare codons at 5′-ends, in terms of reducing (not increasing) RNA folding stability to enable translation initiation (Bentele et al. 2013). Combined with our analysis, the inference that rare codons and positive charges are enriched at 5′-ends/N-termini to enable ribosome slowing now seems an unparsimonious model.

An immediate problem for the ramp hypothesis is our observation that any excess ribosomal occupancy is seen only at the very start of transcripts in *E. coli*. Why might this be? The ramp is defined as a net increase in mean ribosomal occupancy as one moves toward the 5′-end of transcripts. Recently, it has been suggested that high initiation rates on shorter transcripts will give a higher mean occupancy at 5′-ends when averaged over multiple transcripts of all lengths, but need not necessarily be seen in any given transcript (Shah et al. 2013). In principle, if initiation rates are not biased toward small transcripts in *E. coli*, such a statistical artifact could explain the apparent species differences (fig. 8). However, our analysis is normalized by mean transcript occupancy before averaging across genes. As we see on average a downward trend in yeast (fig. 8), a factor other than, or in addition to, short transcripts undergoing more frequent

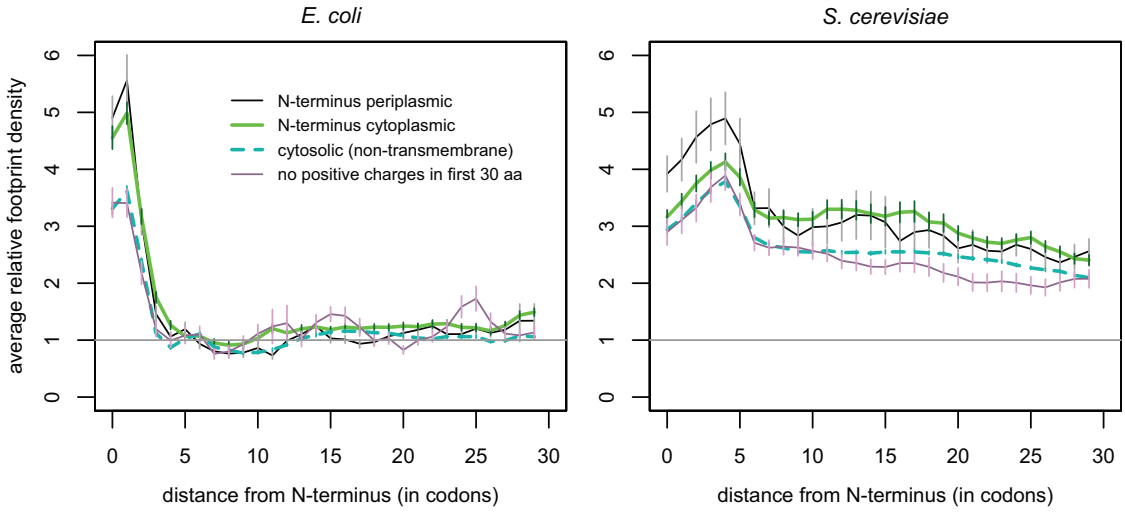

Fig. 8. Relative ribosomal occupancy at the beginning of transcripts. Average within-transcript relative ribosomal occupancies were calculated for different subsets of genes as described in Materials and Methods, "Ribosomal footprint data." In short, the *y* axis represents the changes in ribosomal occupancy from one codon position to the next relative to the occupancy average per site of that transcript, these relative values then being averaged over aligned transcripts. The line at $y = 1$ represents the point at which the ribosomal occupancy in a given position is equal to the average ribosomal occupancy per site of that gene. In all plots, the most increased ribosomal occupancy is seen at the start (approximately the first 4–6 codons or 12–18 nucleotides) of transcripts. *Escherichia coli*: Excess occupancy is seen in all categories, but particularly among transmembrane proteins, but strictly only for the first ~4 codons of a transcript. That the relative ribosomal densities return to the ribosomal occupancy average ($y = 1$) after just a few codons, for all protein categories, strongly suggests this initial ribosomal excess is an initiation artifact (see also Discussion). For the rest of the gene ($x > 4$), standardized major axis regression test that $y \sim x$ slope is not different from 0 from $4 > x < 30$, $P < 2.2e{-}16$ with positive (i.e., increasing approaching $x = 30$) slopes given for all plotted categories, contradicting the downward slope that a ramp would predict. *Saccharomyces cerevisiae*: Excess occupancy is, somewhat similarly to *E. coli*, particularly enriched in all categories at the extreme 5′-end (up to about $x = 6$), but even after this, occupancy is visibly enriched above the gene average and continues to decrease along the length of the plot.

initiation may be required to explain all of the observed 5′ ribosomal densities in this organism.

A further issue for the ramp hypothesis is whether the high ribosomal occupancy seen in both *E. coli* and yeast in close proximity to the start codon (~4 codons in *E. coli*, ~6 in yeast; fig. 8) need reflect elongating ribosomes, as the ramp model presumes. In both eukaryotes and prokaryotes, it is possible for small subunits that have not yet bound a large subunit or completed initiation to bind the start site (Kozak 1999). It may be possible that such noninitiated small subunits are detected by general endonuclease footprinting protocols. Indeed, similar to our results (fig. 8), other footprinting data sets in *E. coli* (Oh et al. 2011) and a human embryonic kidney cell line (Reid and Nicchitta 2012) also profiled short spikes in ribosomal density over only a few codons at the most 5′-ends of transcripts. These short occluded distances at the extreme 5′-end are roughly consistent with the 16–17 nt, which are occluded at the start of the coding sequence by a small subunit at the start codon in yeast (Anthony and Merrick 1992). We suggest that the differences in elevated average relative ribosomal densities (along the first few codons at least) in all plotted protein subgroups may to some extent simply reflect differences in translation initiation rates.

The above interpretation may be consistent with apparently different results, dependent on method (Ingolia et al. 2011). For example, addition of the nonhydrolyzable guanosine triphosphate (GTP) analog Guanylyl imidodiphosphate (GMP-PNP) prevents full (GTP-dependent) ribosome initiation complex formation, leading to an accumulation of small ribosomal subunits positioned at start codons which occlude about 16–17 nt of the start of the coding sequence (Anthony and Merrick 1992). The use of an initiation inhibitor in making the *E. coli* data set under consideration (Li et al. 2012) could then potentially exacerbate the problem of enriched footprints in this region that in fact correspond to nonelongating ribosomes. Additionally, the 5′ charge density may also arise from the fact that GMP-PNP should not have an effect on already-formed initiation complexes that are ready to immediately start translating. Such preformed complexes might be able to translate a couple of codons before the elongation inhibitor chloramphenicol (Li et al. 2012) or cyclohexamide (Ingolia et al. 2009) are able to act. Such a mechanism is consistent with the slight upswing in ribosomal density in the 1–2 codons in *E. coli* or 4 codons in *S. cerevisiae* after the translational start (fig. 8). It is also consistent with the wider initial ribosomal excess (over ~6 codons) in yeast compared with *E. coli*, which might result from fewer codons being strictly stalled over the start codon by the initiation inhibitor.

We must emphasize that we do not wish to claim it is necessarily the case that at least some of the increased ribosomal density at 5′ transcript ends is an artifact of the method used to suppress translation, merely that it is a possibility. There might be true taxon-specific differences in initiation or elongation that cause the observed differences in 5′

ribosomal densities in either *E. coli* or yeast, or both. Nor has the potentially confounding issue of ribosomal drop off yet been addressed. Understanding to what extent a 5′ excess of ribosomes is a result of true taxonomic differences versus a methodological and/or statistical artifact must be a high priority. It remains to be discovered whether positive charges have been under selection to modulate ribosome velocity.

Materials and Methods

Sequences

The June 2008 release of *Saccharomyces* Genome Database gene sequences was obtained from the eukaryotic University of California Santa Cruz (UCSC) Table Browser (Karolchik et al. 2004) at http://genome.ucsc.edu/ (last accessed October 16, 2013) and most other eukaryotic coding sequences were obtained from http://genome.ucsc.edu/ (last accessed October 16, 2013) on 6 April 2013. However, to facilitate analysis of ribosomal footprinting data by Ingolia et al., annotations of the *S. cerevisiae* S288C genome as available on 22 June 2008 (the build used by Ingolia et al. 2009) were obtained separately from the *S. cerevisiae* Genome Database (www.yeastgenome.org, last accessed October 16, 2013); only protein-coding sequences of nondubious classification were considered. Archaeal RefSeq (and the bacterial Refseqs used for making fig. 1) nucleotide sequences coding for protein were downloaded from the microbial UCSC table browser via http://microbes.ucsc.edu/ (last accessed October 16, 2013) on 26 March 2011. For bacterial genomes, we considered the latest release of the European Molecular Biology Laboratory (EMBL) bacterial genome set (http://www.ebi.ac.uk/genomes/bacteria.txt, last accessed October 16, 2013), downloading each genome from the European Bioinformatics Institute (EBI) using a purpose written web crawler. We then considered one genome from each bacterial genus.

For all organisms, any sequences containing nonsense codons or which were not multiples of three were excluded. The sequences were further filtered to only allow the standard or alternative start codons indicated in the appropriate NCBI genetic code table from http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi (last accessed October 16, 2013). Table 1 was used for eukaryotes, table 4 for mycoplasmas, and table 11 for other bacteria as well as archaea. The remaining nucleic acid sequences were translated into protein according to these tables.

Protein Localizations

Escherichia coli cytoplasmic (and other) proteins for figure 2 were obtained from supplementary table S5 of Han et al. (2011). Only those proteins that had all forms of evidence supporting their localization were considered. Although another attempt to sublocalize *E. coli* proteins has been made (Lopez-Campistrous et al. 2005), this data set is much smaller, and there is little overlap in the Swissprot annotations to which the authors compare their localizations, with only 42 proteins agreed to be in the cytosol by both sources. For these reasons, we use the larger data set (Han et al. 2011).

Saccharomyces cerevisiae localizations experimentally determined using a green fluorescent protein reporter construct (Huh et al. 2003) were downloaded from http://yeastgfp.yeastgenome.org (last accessed October 16, 2013). Proteins were allowed to localize to more than one location.

Protein Topology

The current release of TOPDB, a membrane protein topology database, which is based on experimental structural and topological information (Tusnady et al. 2008), was downloaded in xml format from http://topdb.enzim.hu (last accessed October 16, 2013) on 25 March 2013. We limit our analyses using information from this database to *E. coli* due to need of a sufficient sample size of transmembrane proteins within an organism. N-Periplasmic peptide signals were taken from the relevant TOPDB annotations.

Membrane protein topologies based on experimental protein fusions for *S. cerevisiae* were taken from supplementary table S2 of Kim et al. (2006). All C-termini in this table were incorporated in our analysis as they all have direct experimental evidence supporting their topology. Those N-termini with topologies supported by both hidden Markov models (HMMs) were considered in the main text. Yeast proteins with signal peptides were downloaded from the Ensembl Biomart data set EF4 at http://www.ensembl.org/biomart/martview (last accessed October 16, 2013).

Calculating the Average Proportion of Positive Charge in a Given Site across Proteins

To calculate the tendency for positive charge to be used at a certain distance from the N- or C-terminus within a given species, the set of proteins under consideration were aligned by their N- or C-, as appropriate, termini. The amino acids arginine, lysine, and histidine were assigned a charge of 1 and all other amino acids were assigned as 0 (not positively charged). The average proportion of positive charge was then calculated in aligned positions. In all analyses, the first amino acid at the N-terminus is ignored because it is always uncharged. If a protein is less than 60 amino acids in length, only half of the residues within that protein were considered to prevent interference of selection on charge at the opposite terminus. All plots consider the protein terminus (C- or N-, as appropriate) to be at $x = 0$.

Determination of Increasing Average Positive Charge at N-Termini

Some plots of the average proportion of positive charge along the aligned 30 most N-terminal amino acids are best fitted by higher-order equations (as determined by analysis of variance [ANOVA] of nested models). To provide a statistic for whether average positive charge usage increases nearing N-termini in these cases, we note that for an equation of the form $y = ax^n + bx + c$, the slope at any point on the curve is given by $dy/dx = n \times ax^{n-1} + b$. Thus, at the extreme ($x = 0$), regardless of the order of the regression, $dy/dx = b$. This means if the linear term coefficient b is negative, we infer the use of positive charge increases approaching the

N-terminus, and the strength of the increase is reflected in the magnitude of *b*. Linear models and other statistical analyses generally were done in R (R Development Core Team 2010).

Robustness of Increasing or Decreasing Positive Charge Patterns at Termini to Topology Annotation

Our inference that the positive charge usage at N-termini results from the cytosolic orientation of the N-terminus relies upon the presumption that the TOPDB database we use does not rely on the use of positive charge to assign topologies. To this we note the annotations in the TOPDB database are based on experimental evidence, both structural and topological (Tusnady et al. 2008). In combination with this information, the database uses an HMM algorithm (Tusnady and Simon 1998) trained on an experimentally determined, well-defined set of topologies to help predict unknown topologies. Although the HMM considers that different structural parts of a protein (e.g., transmembrane segments, loops) are likely to show an amino acid composition which is divergent compared with the amino acid usage of the protein as a whole, it makes no stipulations about what those amino acid compositions must be. That is, the HMM does not assign membrane topologies by enforcing predetermined rules governing the usage of positive charge, or any other physiochemical property of amino acids, on either side of a transmembrane region.

Nonetheless, we wanted to ensure that the increasing positive charge pattern we detect at cytosolic N-termini is not the result of positive charges in the N-termini of the training set being propagated (or erroneously propagated) through to the topology prediction of N-termini. We find that our results in *E. coli* are indeed robust to using topology annotations supported by increasing levels of experimental evidence, including experimental evidence gathered at the N-terminus specifically (supplementary fig. S7, Supplementary Material online).

We also note that all the HMM-predicted N-termini topologies in yeast (Kim et al. 2006) are constrained by experimental information regarding the topology of the C-terminus that the authors produced in the same article. One of the HMMs used to predict the N-terminal topologies, prodiv-TMHMM, relies on a similar method to the one used by TOPDB, and does not explicitly use positive charge to infer the most likely membrane orientation (Viklund and Elofsson 2004). The other HMM employed by Kim et al., TMHMM, does incorporate (among other factors) charge bias in its determinations of membrane protein topology (Krogh et al. 2001). In the main text, we use those proteins whose topologies are supported by both of these two independent methods. For completeness and transparency, we have also examined the increase in positive charge use among those proteins topologies predicted by each HMM separately. We can report that these additional analyses give similar results to those presented in the main text, namely that increasing N-terminal positive charge is observed only among those membrane proteins whose N-termini reside in the cytosol (and in the absence of signal sequences) (supplementary fig. S8, Supplementary Material online).

Determination of Positions of Significant Average Positive Charge Enrichment at N-Termini

Within a group of proteins for which we perform a regression of positive charge usage on distance from N-terminus, we determined the location(s) of significantly increased average positive charge by 1,000 iterations of the following method. The sequences of all proteins within the considered group were shuffled, and the proportion of positive charge within the randomized sequences was calculated in each of the first 30 positions near the N-terminus. For each iteration, if the proportion of randomized positive charge in a given position is greater than or equal to the proportion of positive charge observed in that position within the considered group, m is incremented in that position. P for each position is then calculated as $(m + 1)/(n + 1)$, where n is the number of iterations performed.

Determination of the Point of Maximal or Minimal Positive Charge Usage

For second-order equations and higher, the point of maximal or minimal charge usage corresponds to the point where the slope of the tangent is zero. This point was determined by setting the derivative of each linear model equal to zero and solving the equations in MATLAB (2010).

Reconstruction of N-Terminal Positive Charge According to Trends in Charge Usage near Transmembrane Regions

If proteins transitioning from the cytosol into membranes can indeed account for the increasing positive charge usage at the N-terminus, we should be able to reproduce the observed pattern of increasing average positive charge approaching cytosolic N-termini given solely the locations of where these cytosol-to-membrane transitions occur. To measure trends in positive charge usage near membranes outside of any possible additional selection on positive charge within the first 30 amino acids, we consider in this analysis only those transmembrane regions within proteins where the cytosolic N-terminus transitions into the membrane at least 45 amino acids downstream of the start of the protein. For each protein with such a region, we recorded the number of positive charges used in each of six consecutive windows, each five amino acids in length, directly surrounding the cytosolic face of the membrane such that the first three windows cover cytosolic amino acids and the latter three windows cover amino acids situated in the membrane. The average number of positive charges in each window (defined by its location relative to the membrane) was then calculated across all suitable proteins.

We then returned to the distribution of locations where cytoplasmic N-termini come into contact with membranes within the first 30 amino acids at the N-terminus. The positive charge at each of (up to) 30 amino acid positions surrounding the N-terminal point of transition into the membrane was

incremented by the density of positive charges within the analogous observed window as calculated above. The average number of positive charges, as reconstructed, at each position was then calculated.

Ribosomal Footprint Data

Ribosomal densities derived from two replicates of ribosomally protected fragments along the transcriptome of *E. coli* (Li et al. 2012) were downloaded from GSM872393 and GSM872394 at http://www.ncbi.nlm.nih.gov/geo/ (last accessed October 16, 2013). Positions in each replicate that had no footprint counts available were given a footprint count of zero. The ribosomal footprint counts at each position along the transcriptome were averaged between the two replicates.

Sequenced ribosomally protected fragments for *S. cerevisiae* grown in rich media, data set GSE13750 (Ingolia et al. 2009) were downloaded from the NCBI Gene Expression Omnibus at www.ncbi.nlm.nih.gov/projects/geo (last accessed October 16, 2013). Only one mismatch between the sequenced fragment and reference genome sequence was allowed. The chromosomal location and coordinates of the sequenced fragments given in the original data set were used in combination with the start and stop coordinates from the gene annotations (see Materials and Methods, sequences) to map footprints to transcripts. All fragment counts were taken as the average value of the two experimental replicates and only footprints that mapped uniquely to one location in the reference genome were considered. In line with Ingolia et al., we assigned footprints to genes if the first base of the footprint mapped to 16 nt before the first base or 14 nt before the last base of the gene, to take account of which area of the footprint is likely in the ribosomal active site.

We consider that whether ribosomal occupancy is, on average, enriched at transcript starts is best addressed by normalizing the ribosomal density at the start of a given transcript relative to the average ribosomal occupancy of the same transcript. This has the advantage of treating footprints as increases or decreases in density within the context of a single transcript (a ribosome travels only along a single mRNA at a time) and sidesteps the ambiguity in deciphering emergent patterns that might result from raw footprint counts being averaged across different transcripts. To this end, we calculated relative ribosomal occupancies for the first 30 codons ($1 \geq x \leq 30$) in each *E. coli* transcript by dividing the ribosomal density at position x by the average ribosomal occupancy of the entire gene. Transcripts were then aligned by their 5′-ends and the mean relative ribosomal occupancy in each position was calculated to create figure 8.

Supplementary Material

Supplementary figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a University Research Studentship from the University of Bath to C.A.C. and the Wolfson Royal Society Research Merit Award to L.D.H.

References

Anthony DD, Merrick WC. 1992. Analysis of 40 S and 80 S complexes with mRNA as measured by sucrose density gradients and primer extension inhibition. *J Biol Chem.* 267:1554–1562.

Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol.* 9:675.

Berezovsky IN, Kilosanidze GT, Tumanyan VG, Kisselev LL. 1999. Amino acid composition of protein termini are biased in different manners. *Protein Eng.* 12:23–30.

Bjornsson A, Mottagui-Tabar S, Isaksson LA. 1996. Structure of the C-terminal end of the nascent peptide influences translation termination. *EMBO J.* 15:1696–1704.

Charneski CA, Hurst LD. 2013. Positively charged residues are the primary determinants of ribosomal velocity. *PLoS Biol.* 11:e1001508.

Dalbey RE, Wang P, Kuhn A. 2011. Assembly of bacterial inner membrane proteins. *Annu Rev Biochem.* 80:161–187.

Delgado-Partin VM, Dalbey RE. 1998. The proton motive force, acting on acidic residues, promotes translocation of amino-terminal domains of membrane proteins when the hydrophobicity of the translocation signal is low. *J Biol Chem.* 273:9927–9934.

de Smit MH, van Duin J. 1990. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci U S A.* 87:7668–7672.

Dimitrova LN, Kuroha K, Tatematsu T, Inada T. 2009. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *J Biol Chem.* 284:10343–10352.

Eyre-Walker A, Bulmer M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.* 21:4599–4603.

Gafvelin G, Sakaguchi M, Andersson H, von Heijne G. 1997. Topological rules for membrane protein assembly in eukaryotic cells. *J Biol Chem.* 272:6119–6127.

Gallusser A, Kuhn A. 1990. Initial steps in protein membrane insertion. Bacteriophage M13 procoat protein binds to the membrane surface by electrostatic interaction. *EMBO J.* 9:2723–2729.

Goder V, Junne T, Spiess M. 2004. Sec61p contributes to signal sequence orientation according to the positive-inside rule. *Mol Biol Cell.* 15:1470–1478.

Han MJ, Yun H, Lee JW, Lee YH, Lee SY, Yoo JS, Kim JY, Kim JF, Hur CG. 2011. Genome-wide identification of the subcellular localization of the *Escherichia coli* B proteome using experimental and computational methods. *Proteomics* 11:1213–1227.

Heijne G. 1986. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* 5:3021–3027.

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. *Nature* 425:686–691.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218–223.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147:789–802.

Islam MM, Khan MA, Kuroda Y. 2012. Analysis of amino acid contributions to protein solubility using short peptide tags fused to a simplified BPTI variant. *Biochim Biophys Acta.* 1824:1144–1150.

Ito-Harashima S, Kuroha K, Tatematsu T, Inada T. 2007. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Genes Dev.* 21:519–524.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* 32:D493–D496.

Kiefer D, Hu X, Dalbey R, Kuhn A. 1997. Negatively charged amino acid residues play an active role in orienting the sec-independent Pf3 coat protein in the *Escherichia coli* inner membrane. *EMBO J.* 16: 2197–2204.

Kim H, Melen K, Osterberg M, von Heijne G. 2006. A global topology map of the *Saccharomyces cerevisiae* membrane proteome. *Proc Natl Acad Sci U S A.* 103:11142–11147.

Kozak M. 1999. Initiation of translation in prokaryotes and eukaryotes. *Gene* 234:187–208.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 305:567–580.

Kuroiwa T, Sakaguchi M, Mihara K, Omura T. 1990. Structural requirements for interruption of protein translocation across rough endoplasmic reticulum membrane. *J Biochem.* 108:829–834.

Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484:538–541.

Li P, Beckwith J, Inouye H. 1988. Alteration of the amino terminus of the mature sequence of a periplasmic protein can severely affect protein export in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 85:7685–7689.

Lopez-Campistrous A, Semchuk P, Burke L, Palmer-Stone T, Brokx SJ, Broderick G, Bottorff D, Bolch S, Weiner JH, Ellison MJ. 2005. Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth. *Mol Cell Proteomics.* 4:1205–1209.

Lu J, Deutsch C. 2008. Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J Mol Biol.* 384:73–86.

Lu J, Kobertz WR, Deutsch C. 2007. Mapping the electrostatic potential within the ribosomal exit tunnel. *J Mol Biol.* 371:1378–1391.

MATLAB. 2010. Version 7.11.0 (R2010b). Natick (MA): The MathWorks Inc.

McCarthy JE, Bokelmann C. 1988. Determinants of translational initiation efficiency in the atp operon of *Escherichia coli*. *Mol Microbiol.* 2: 455–465.

Mitarai N, Sneppen K, Pedersen S. 2008. Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. *J Mol Biol.* 382:236–245.

Moarefi I, Jeruzalmi D, Turner J, O'Donnell M, Kuriyan J. 2000. Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. *J Mol Biol.* 296:1215–1223.

Mottagui-Tabar S, Bjornsson A, Isaksson LA. 1994. The second to last amino acid in the nascent peptide as a codon context determinant. *EMBO J.* 13:249–257.

Nilsson I, von Heijne G. 1990. Fine-tuning the topology of a polytopic membrane protein: role of positively and negatively charged amino acids. *Cell* 62:1135–1141.

Nishiyama K, Maeda M, Yanagisawa K, Nagase R, Komura H, Iwashita T, Yamagaki T, Kusumoto S, Tokuda H, Shimamoto K. 2012. MPIase is a glycolipozyme essential for membrane protein integration. *Nat Commun.* 3:1260.

Oh E, Becker AH, Sandikci A, et al. (12 co-authors). 2011. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* 147:1295–1308.

Puziss JW, Fikes JD, Bassford PJ Jr. 1989. Analysis of mutational alterations in the hydrophilic segment of the maltose-binding protein signal peptide. *J Bacteriol.* 171:2303–2311.

Qian W, Yang J-R, Pearson NM, Maclean C, Zhang J. 2012. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.* 8:e1002603.

R Development Core Team. 2010. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Reid DW, Nicchitta CV. 2012. Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. *J Biol Chem.* 287:5518–5527.

Samuelson JC, Chen M, Jiang F, Moller I, Wiedmann M, Kuhn A, Phillips GJ, Dalbey RE. 2000. YidC mediates membrane protein insertion in bacteria. *Nature* 406:637–641.

Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation. *Cell* 153:1589–1601.

Sipos L, von Heijne G. 1993. Predicting the topology of eukaryotic membrane proteins. *Eur J Biochem.* 213:1333–1340.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141:344–354.

Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.* 12:R110.

Tusnady GE, Kalmar L, Simon I. 2008. TOPDB: topology data bank of transmembrane proteins. *Nucleic Acids Res.* 36:D234–D239.

Tusnady GE, Simon I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol.* 283:489–506.

van Klompenburg W, Nilsson I, von Heijne G, de Kruijff B. 1997. Anionic phospholipids are determinants of membrane protein topology. *EMBO J.* 16:4261–4266.

Viklund H, Elofsson A. 2004. Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.* 13:1908–1917.

Vlasuk GP, Inouye S, Ito H, Itakura K, Inouye M. 1983. Effects of the complete removal of basic amino acid residues from the signal peptide on secretion of lipoprotein in *Escherichia coli*. *J Biol Chem.* 258:7141–7148.

von Heijne G. 1984. Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implications for protein export in prokaryotic and eukaryotic cells. *EMBO J.* 3: 2315–2318.

von Heijne G, Gavel Y. 1988. Topogenic signals in integral membrane proteins. *Eur J Biochem.* 174:671–678.

Whitley P, Zander T, Ehrmann M, Haardt M, Bremer E, von Heijne G. 1994. Sec-independent translocation of a 100-residue periplasmic N-terminal tail in the *E. coli* inner membrane protein proW. *EMBO J.* 13:4653–4661.

Yamane K, Mizushima S. 1988. Introduction of basic amino acid residues after the signal peptide inhibits protein translocation across the cytoplasmic membrane of *Escherichia coli*. Relation to the orientation of membrane proteins. *J Biol Chem.* 263: 19690–19696.

V. Atypical AT skew in Firmicute genomes results from selection and not from mutation

Catherine A. Charneski & Laurence D. Hurst
PLoS Genet (2011) 7(9): e1002283

OPEN ACCESS Freely available online PLoS GENETICS

Atypical AT Skew in Firmicute Genomes Results from Selection and Not from Mutation

Catherine A. Charneski[1], Frank Honti[1,2], Josephine M. Bryant[1,3], Laurence D. Hurst[1☯]*, Edward J. Feil[1☯]*

1 Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom, **2** Medical Research Council Functional Genomics Unit, Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, United Kingdom, **3** The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

Abstract

The second parity rule states that, if there is no bias in mutation or selection, then within each strand of DNA complementary bases are present at approximately equal frequencies. In bacteria, however, there is commonly an excess of G (over C) and, to a lesser extent, T (over A) in the replicatory leading strand. The low G+C Firmicutes, such as *Staphylococcus aureus*, are unusual in displaying an excess of A over T on the leading strand. As mutation has been established as a major force in the generation of such skews across various bacterial taxa, this anomaly has been assumed to reflect unusual mutation biases in Firmicute genomes. Here we show that this is not the case and that mutation bias does not explain the atypical AT skew seen in *S. aureus*. First, recently arisen intergenic SNPs predict the classical replication-derived equilibrium enrichment of T relative to A, contrary to what is observed. Second, sites predicted to be under weak purifying selection display only weak AT skew. Third, AT skew is primarily associated with largely non-synonymous first and second codon sites and is seen with respect to their sense direction, not which replicating strand they lie on. The atypical AT skew we show to be a consequence of the strong bias for genes to be co-oriented with the replicating fork, coupled with the selective avoidance of both stop codons and costly amino acids, which tend to have T-rich codons. That intergenic sequence has more A than T, while at mutational equilibrium a preponderance of T is expected, points to a possible further unresolved selective source of skew.

Citation: Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ (2011) Atypical AT Skew in Firmicute Genomes Results from Selection and Not from Mutation. PLoS Genet 7(9): e1002283. doi:10.1371/journal.pgen.1002283

Editor: Dmitri A. Petrov, Stanford University, United States of America

Received February 25, 2011; **Accepted** July 12, 2011; **Published** September 15, 2011

Funding: The work was funded by the University of Bath Overseas Research Studentship. LDH is a Royal Society Wolfson Research Merit Award Holder. EJF is funded by the TROCAR consortium (EU FP7-HEALTH #223031), http://www.trocarproject.eu/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: l.d.hurst@bath.ac.uk (LDH); e.feil@bath.ac.uk (EJF)

☯ These authors contributed equally to this work.

Introduction

Skews in nucleotide usage (compositional asymmetries) are of interest as they provide a window into fundamental processes operating within genomes. Under conditions of equal mutation bias and random gene orientation, the two complementary strands of a bacterial chromosome should be subject to the same sets of substitutions, and hence each should contain approximately equal amounts of a given base and its complement [1]. This condition, where A~T and C~G within a given strand, is known as the second parity rule and represents a null expectation of sequence evolution. The division of the replication fork into leading and lagging strands, however, has shaped bacterial sequence evolution contrary to this null, as each strand generally possesses an excess of one nucleotide over its complementary base (called GC and AT skews). Within bacterial genomes, nucleotide skews normally manifest as a richness of G over C and (with a lesser magnitude) T over A on the replicatory leading strand [2–5].

These genomic skews indicate some force, be it mutation or selection, is biasing substitutions between the two replicating strands. While it is acknowledged that, in theory, selection for genes to reside in the leading strand coupled with preferences for particular amino acids could result in chromosome-wide skews [2,5–8], such a role for selection in generating large-scale compositional bias remains largely hypothetical and undescribed. Instead mutational biases between the two replicating strands are generally invoked as the cause of nucleotide skew [3,8,9]. Mutational differences between transcribed and non-transcribed strands have also been considered [10,11], and these explanations incorporate a selective element as they require asymmetrically distributed genes between the replicating strands.

It has been argued that strand-specific mutation biases might result from the different amounts of time spent by each strand exposed in the single-stranded state during continuous or discontinuous DNA replication. While cytosine deamination (C→T) in particular was long suspected to play a major role in creating the excess of G and T in the leading strand, it has been shown that similar compositional skews can result from a variety of mutational scenarios [12]. The observation that GC skews tend to be stronger than AT skews also points to contributions from multiple mutation types. As would be expected if they are primarily mutational in origin, detected skews are generally higher in nearly neutral sites such as intergenic regions and fourfold degenerate sites [2,3,10].

Author Summary

When considering a single strand of DNA, it is not necessarily the case that the frequency of each base should equal its complementary partner, such that A = T and G = C. For the leading strand, it is typically the case that Gs are more common than Cs, and Ts more common than As. This bias is widely thought to arise due to different mutational biases during replication. The Firmicutes exhibit an atypical preference for A over T on the leading strand, and here we show that selection, rather than mutation, can explain this exception. For those bases within coding regions, selection acts to inflate the frequency of A over T in order to avoid stop codons and to use metabolically cheap amino acids. Because genes are not orientated randomly, this manifests as an overall enrichment of A on the leading strand. Furthermore, a direct examination of mutational patterns is inconsistent with the observed enrichment of As. Curiously, our data also point to an unresolved source of selection on synonymous and intergenic sites, which are widely assumed to be neutral.

Staphylococcus aureus is an unusual case in that, like other Firmicutes, it displays an excess of A over T in the leading strand, or positive AT skew given as (A–T)/(A+T) [13]. Why does this AT skew run counter to that observed in most bacteria? One possibility is that unique selective processes might be avoiding T and preferring A in the leading strand. Genes predominately lie in the leading strand in *S. aureus*, a feature of bacterial chromosomes posited to result from selection to minimize impacts between DNA and RNA polymerases [14] (although the relevance of this mechanism remains unclear). Any pressure to underuse codons rich in T could then result in AT skews simply due to the differential coding content of these two strands. Gene orientation bias is particularly enhanced in low G+C Firmicutes, potentially on account of the replication fork asymmetry induced by the possession of separate α subunits for synthesis of the leading and lagging strands [15,16]. Alternatively, *S. aureus* might display a mutational bias which produces AT skew opposite that of most other bacteria, pushing up A over T in the leading strand. Indeed, it was recently suggested that the DNA polymerase-α subunit that replicates the leading strand also determines the direction of AT skew [16]. However, this finding was not repeated in a subsequent study and a direct mutational effect on AT skew resulting from α-subunit possession was called into question [11].

Here we investigate whether mutation or selection best explains the unusual AT skew in *S. aureus*. Dividing the chromosome into coding and non-coding positions allowed us to assess whether skew is strongest in those sites which should be under weaker purifying selection, such as intergenic and fourfold degenerate sites, or whether skew is most prevalent in non-synonymous sites which are constrained by the need to code for amino acids. Moreover we make use of newly described, high resolution genome-wide SNP data representing a single widespread clone of methicillin-resistant *Staphylococcus aureus* (MRSA) [17]. As these isolates have diverged from a very recent common ancestor, over a period of 4-5 decades, the data provide an opportunity to infer mutational patterns in *S. aureus* and contrast AT skews expected under mutational equilibrium to the AT skews observed. Importantly, the false positive rate of SNP calling in these genomes is benchmarked to be less than 1 SNP per genome (Julian Parkhill, personal communication), making these an unprecedentedly high quality resource.

Results

AT skew in *S. aureus* is unusual

As some of our results focus on coding sites within a single DNA molecule, the published strand, whereas others utilize coding sites in the sense direction on either the leading or lagging strands, we have provided a schematic to illustrate which sites are being considered in different types of analyses (Figure 1).

A plot of AT skew on the published strand in non-overlapping windows confirms that AT skew in *S. aureus* is unlike that of most bacteria as it is positive in the first half of the published strand (Figure 2), as previously described [13]. Considering only the core (vertically transmitted) without non-core (laterally transferred) regions eliminates irregularities in the AT skew which may arise from the importation of sequences which previously resided on an oppositely-skewed strand (Figure 2). For this reason the core genome only was considered in the rest of our analysis.

AT skew in *S. aureus* is not primarily mutational

Three lines of evidence argue against mutation as the cause of AT skew in *S. aureus*:

1) Evidence from comparison of weakly and more strongly selected sites. If skew were primarily mutational, we should expect to see the strongest skew associated with sites which are under comparatively weak selective constraint. Separating the genomic skew into different sites relative to the published strand reveals there is, at best, only a small mutagenic contribution to AT skew evident in weakly-selected intergenic regions and fourfold degenerate sites, where mutations are synonymous (Figure 3, Table 1). Instead the greatest AT skew, by an order of magnitude, is seen in first and second codon positions (Figure 3, Table 1), sites which are largely non-synonymous and should be more buffered against the effects of mutational biases.

That fourfold degenerate sites show only a small contribution to overall AT bias provides no evidence for transcription-associated mutation as the prime cause of skew. Even in the case that codon usage bias is acting to alter the transcriptional effects on AT skew in fourfold sites, we still observe that intra-operonic intergenic regions, which are putatively transcribed, display a very weak AT skew similar to that seen in ex-operonic intergenic regions and fourfold sites (Table 1). We thus conclude that if there is a transcriptional mutation bias affecting AT skew, it is rather slight. Indeed, if anything, the lower intra-operonic leading-strand AT skew values relative to ex-operonic leading-strand AT skew is consistent with a transcriptional pressure towards T (over A).

Additionally, if transcriptional mutation were the cause of atypical AT skew, we should expect to observe a decrease in skew with increasing distance from gene boundaries. This effect would due to the transcription of UTRs of varying lengths by RNA polymerase, with transcription-induced skew decreasing away from gene boundaries as the contribution of increasingly longer UTRs to intergenic regions declines. We would also expect to observe increased AT skew in intra-operonic intergenic sites, which should be more prone to transcriptional effects. Figure 4 shows the patterns of AT skew with increasing distance from 5′ (upstream) and 3′ (downstream) gene boundaries in intra-operonic and ex-operonic sites. Although we note striking deviations in AT skew at both boundaries [see also 18,19], these patterns are not monotonic and therefore cannot be explained by transcriptional effects. Instead, we consider it likely that these deviations correspond to translational initiation and termination signals, and similar effects are observed in other species (Figure S1). Furthermore, the patterns are very similar for intra-operonic and

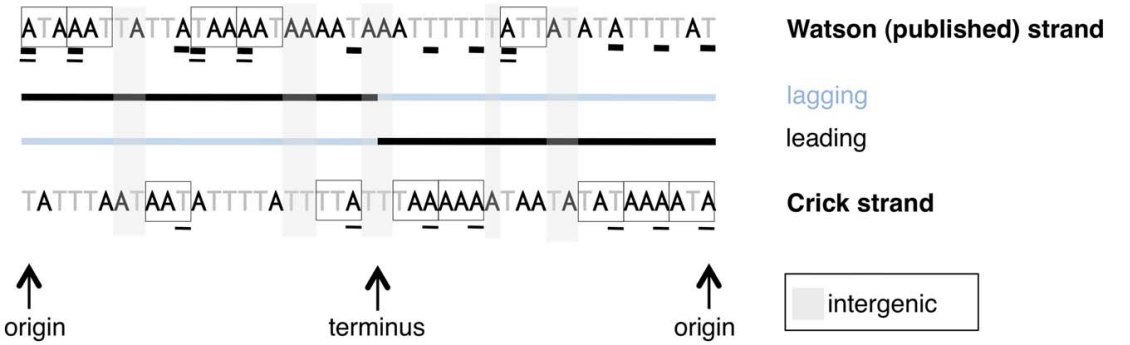

Figure 1. A schematic of sites used for different types of analyses in this paper. A mock linearized *S. aureus* genome is shown, with codons on the strand from which they are transcribed shown in boxes. Note coding content is over-represented on the leading strand. Only first codon positions are considered for the sake of simplicity. The results presented in Figure 2, Figure 3, and Figure 5 consider coding sites underlined with a thick line, meaning the identities of the nucleotides are all taken from the published strand in dedicated coding sites, regardless of the sense direction of the gene. Thus Figure 2 and Figure 3 do not explicitly distinguish between leading and lagging AT skews, but give an averaged picture of skews along both strands of the genome. All other analyses of AT skew in coding sites in this paper consider the sites underlined with a thin line. The identities of these nucleotides are all in the sense direction of the gene, i.e. that nucleotide which appears within the transcript, and may be easily divided into groups according to whether they are encoded on the leading or lagging strand. The analysis of intergenic regions is simpler in concept as intergenic sites clearly divide into either leading or lagging.
doi:10.1371/journal.pgen.1002283.g001

ex-operonic intergenic regions, the increased scatter in the former sites reflecting a smaller sample size (Figure 4). On account of these deviations in skew at the boundaries of intergenic regions, 60 bp were removed from each end of intergenic regions for all subsequent analyses.

2) Evidence from rare SNPs. From examination of rare and putatively weakly-selected SNPs, which should better reflect mutational pressures, we can infer the mutational profile in *S. aureus* and the corresponding nucleotide frequencies expected at mutational equilibrium and hence the expected skew at mutational equilibrium (Table 2). We considered that not only might selection on codon usage be biasing observed SNPs in fourfold degenerate sites, but that fourfold sites are biased in terms of the possible nucleotides that may precede them in the second codon position and thus may give a distorted representation of mutational processes due to over- or under-represented dinucleotide effects. Instead we consider that the mutational profile in intergenic regions would best reflect the AT skew expected to result from replication-associated mutation. Under the premise that intra-operonic intergenic regions are more likely to be subject to

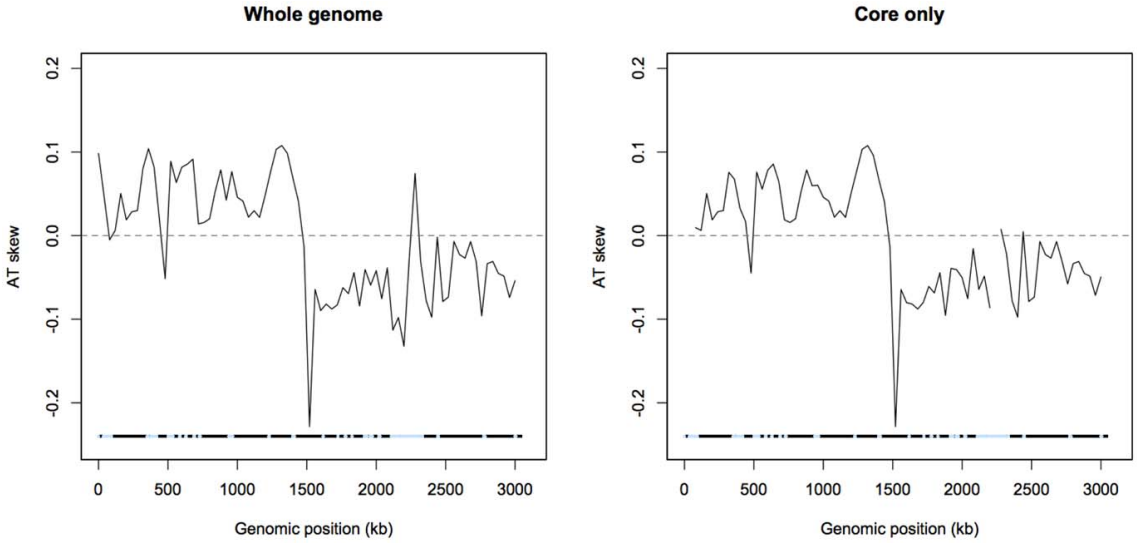

Figure 2. AT skew in *S. aureus*. Skew calculated in 40 kb non-overlapping windows is shown with respect to the published strand. The origin of replication is near 0 kb and the terminus of replication is expected to coincide at the midpoint where the skew changes direction. Excluding non-core (denoted by blue regions) eliminates an A-rich peak at approximately 2,300 kb into the genome.
doi:10.1371/journal.pgen.1002283.g002

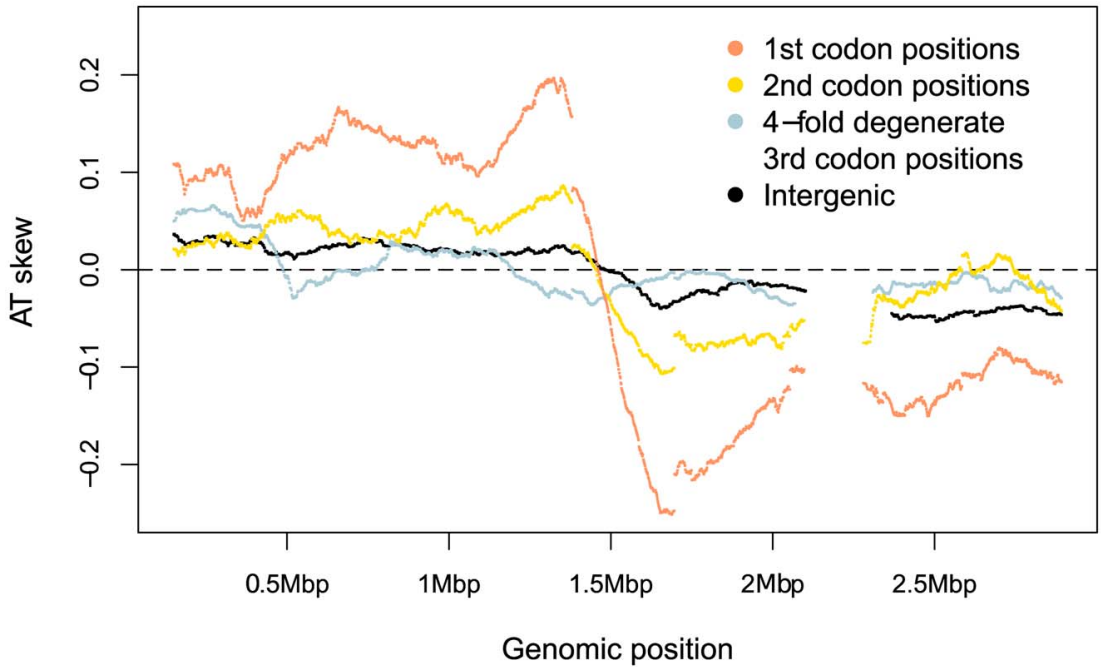

Figure 3. *S. aureus* AT skews in different sites along the genome. Skew was calculated with respect to the published strand using overlapping windows of 300kb and 1 kb steps where a data point was plotted if the number of bases in a window reached at least 30,000. Codon positions were demarcated along the published (Watson) strand and skews in these positions were plotted using the sequence in this single DNA molecule without respect to whether the gene is encoded on the leading or lagging strand.
doi:10.1371/journal.pgen.1002283.g003

transcriptional pressures (e.g. for regulation of translational coupling [20]), we limited our analysis of replication-associated mutation to ex-operonic intergenic regions, intergenic being defined as more than 60 bp from either end of a gene and with a 500 bp maximum length cutoff.

Importantly, ex-operonic intergenic SNPs indicate that single base mutations acting alone would lead to a strong excess of T over A in the intergenic leading strand at compositional equilibrium (Table 2). Bootstrapping the *S. aureus* data (see Methods) supports the view that the expected compositional equilibrium is negative (Table 2). Additionally, the intergenic equilibrium AT skew expected for a second Gram positive Firmicute which also displays an unusual (positive) AT skew in the leading strand, *Bacillus anthracis*, is negative (Table S1). In this case the appropriateness of the SNPs for this sort of analysis is not so easily demonstrated as the sequencing was performed by multiple groups and in several instances we cannot find statements of the quality of the sequencing. However, in a further non-Firmicute for which high quality recent SNPs are available, the Gram negative *Salmonella enterica str. typhi*, we again find that the SNP profile predicts the typical (negative) leading AT skew (Table S1), suggesting the predicted neutral equilibrium T>A bias in the leading strand of *S. aureus* is not unusual. In the case of both *B. anthracis* and *S. typhi*, as the sample size of recent SNPs is relatively small, less confidence can be given to the equilibrium values than in the case of *S. aureus* (Table S1).

Table 1. A quantification of *S. aureus* AT skews.

Site	Leading AT skew	Lagging AT skew	P
1st positions	0.2403	0.2081	<0.0001
2nd positions	0.0717	0.0245	<0.0001
4-fold degenerate sites	0.0199	-0.0033	<0.0001
Intergenic (intra-operonic)	0.0021	-0.0193	0.280
Intergenic (ex-operonic)	0.0276	-0.0276	<0.0001

AT skew values in coding positions in this table are given with respect to the protein-coding sense direction of a gene and are differentiated into whether they lie on the leading or lagging strand (see Figure 1). While ex-operonic skews are calculated from all available ex-operonic sequence, intra-operonic skews are calculated only for the strand on which the surrounding genes are transcribed. P was calculated as $(r+1)/(n+1)$, where r is the number of simulated genomes resulting in a difference in AT skew values between the leading and lagging strands equal to or greater in magnitude than that observed in the reference (TW20) genome, and n is the total number of simulations performed (10,000).
doi:10.1371/journal.pgen.1002283.t001

In addition, if selection were operating on intergenic SNPs we should expect that older SNPs will predict an equilibrium skew closer to that observed, as selection will have had longer to operate on older SNPs, potentially removing the weakly deleterious ones. To test this we examine 54 SNPs that were found in two, three or four *S. aureus* isolates. We find that the predicted equilibrium AT skew obtained from the 54 ex-operonic SNPs present in multiple isolates (equilibrium AT skew = 0.4757, 95% bootstrap interval: 0.0763, 0.8087) is of completely different sign than that obtained using 140 ex-operonic singletons (-0.4176, 95% bootstrap interval: -0.6792, -0.1522). A randomization was used to put a significance level on whether these values are significantly different. The

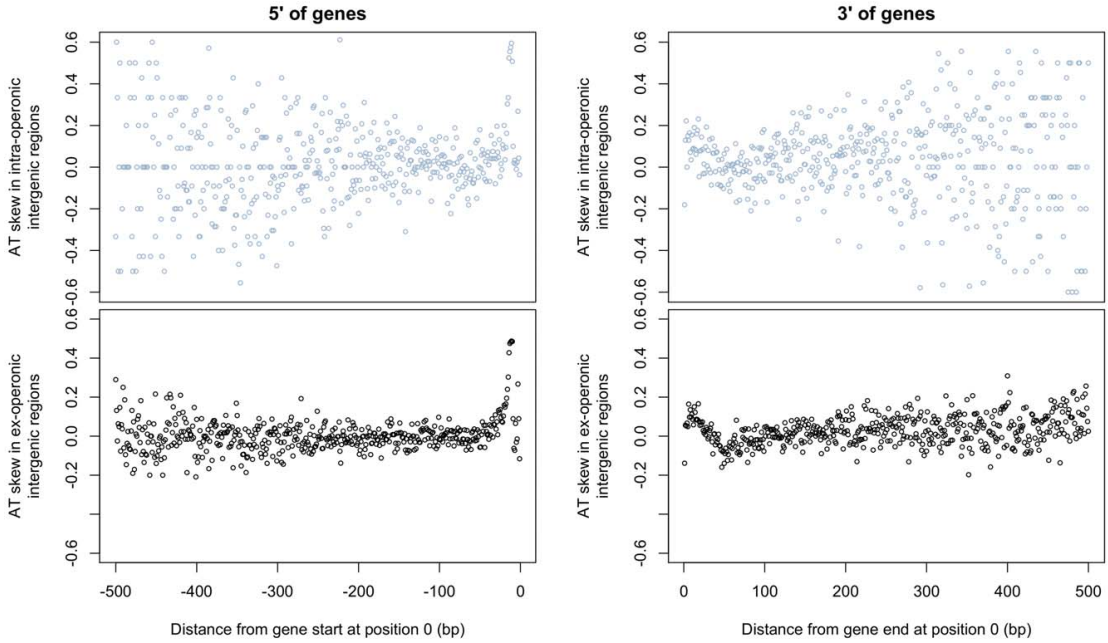

Figure 4. AT skew displays local abnormalities at intergene boundaries but grows neither A nor T rich at increasingly distant positions from gene starts and ends. AT skew at each position was calculated from the nucleotide content measured across all intra- or ex-operonic intergenic regions at that position relative to the gene start or end as appropriate. All intergenic regions were considered in the direction of transcription of the relevant gene, and similar results are obtained when only UTRs of leading strand genes are considered (latter not shown). The effects on AT skew are similar between ex-operonic and intra-operonic intergenic regions for regions both 5′ and 3′ of genes, with more noise apparent at further distances in the intra-operonic regions due to increasingly smaller sample sizes. *5′ of genes*. With increasing distance 5′ of gene starts, AT skew first increases before it decreases. *3′ of genes*. Starting at gene ends, AT skew becomes more positive before it decreases again, becoming briefly negative before levelling out.
doi:10.1371/journal.pgen.1002283.g004

singleton SNPs and those SNPs present in 2, 3, or 4 isolates were combined into one large group. From this group SNPs were sampled with replacement and randomly allocated to either the singleton SNP group (140 SNPs) or the 2, 3, or 4 isolate group (54 SNPs). These singleton and 2, 3, 4 strain groups were randomly simulated 2000 times, and for each simulation the equilibrium AT skew was calculated for the two random groups and the difference between the two equilibria determined. P was calculated as $(r+1)/(n+1)$ where r is the number of simulations which produced a difference between the singleton and 2, 3, 4 group greater than or

Table 2. Relative mutation rates of nucleotide i to j per site i for ex-operonic intergenic sites were calculated from singleton SNPs for the two replicatory strands.

		from A	T	C	G	Equilibrium frequency	Equilibrium AT skew
Leading	to **A**	-	1.79965E-04	4.37222E-04	6.39046E-04	0.2257	-0.4176 (-0.6792, -0.1522)
	T	2.34002E-04	-	1.67602E-03	1.17158E-03	0.5493	
	C	6.38189E-05	4.72409E-04	-	5.32538E-05	0.1319	
	G	6.59462E-04	4.49913E-05	0.00	-	0.0931	
Lagging	to **A**	-	2.34003E-04	1.17158E-03	1.67602E-03	0.5493	0.4176 (0.6792, 0.1522)
	T	1.79965E-04	-	6.39046E-04	4.37222E-04	0.2257	
	C	4.49913E-05	6.59462E-04	-	0.00	0.0931	
	G	4.72409E-04	6.38189E-05	5.32538E-05	-	0.1319	

Relative rates were derived from the following leading strand ex-operonic SNP counts, where XY represents a change from nucleotide X to Y: AG 31 GA 12 CG 0 GC 1 GT 22 TA 8 TC 21 TG 2 AC 3 CA 6 AT 11 CT 23. Nucleotide frequencies at compositional equilibrium were derived from the relative mutation rates. Leading and lagging equilibrium AT skews were calculated from equilibrium A and T frequencies. 95% bootstrap intervals are shown in parentheses.
doi:10.1371/journal.pgen.1002283.t002

equal in magnitude to that observed and n is the number of randomizations performed. This test indicates these two groups of SNPs are different as regards the expected AT skew ($P = 0.0065$). The assumption that older SNPs are more prone to selection is supported by analysis of dN/dS ratios for SNPs in genes: the average dN/dS over all pairwise genomic comparisons is 0.69, but that for genomes diverged by fewer than 10 SNPs approaches 1 [21], indicating that recent SNPs have not yet had time to be purged [22] and better reflect the mutational profile.

The above results support the view that the AT skew cannot be explained by mutation bias, but also have broader implications for inferring mutational patterns. The striking difference in the mutational profiles inferred from singleton SNPs and from those SNPs present in multiple isolates points to a class of mutation which, though non-lethal (and therefore observable), is sufficiently deleterious to be purged very rapidly by selection. Thus, in order to make the most reliable inferences of mutational profiles (and hence predicted equilibria) it is necessary to consider extremely recently-emerged singleton SNPs between very closely related genomes, with the caveat that the low number of SNPs per clone be offset by large samples of genomes.

The reliability of sequence data is of the utmost importance when using singletons, and such SNPs have been avoided in the past owing to the possibility of sequencing errors [23–27]. However, we are confident that false positive SNPs cannot account for our results. With a remarkable benchmarked false-positive rate of no more than 1 per genome (Julian Parkhill, personal communication), the maximum number of false positive ex-operonic SNPs in our analysis is (the fraction of intergenic sequence which is ex-operonic) * (the error rate of 1 base per genome) * (the number of genomes used) $= (125042/3043210)*1*62 \approx 2.55$, or about 3. We used randomizations to assess the effect that these potential miscalled SNPs would have on our AT skew calculations. For each simulation, 3 (hypothetical false positives) of the 140 ex-operonic SNPs were removed at random and the equilibrium AT skew resulting from the remaining SNPs was calculated, with 1,000 simulations performed in total. All of the resulting leading strand equilibrium AT skew values are negative and fall between -0.3830 and -0.4910, indicating false positives are not affecting our inference of the sign of the equilibrium AT skew. For these reasons we contend that an analysis of very recent singletons is the best reflection of the mutational profile. To the best of our knowledge the data set we examine is the only one of high enough quality for recently diverged (post 1960 [17]) lineages.

It is possible that that even the intergenic singleton SNPs have been affected by selection and are hence not an unbiased reflection of the mutation profile. If so, the difference between the true mutational equilibrium and the observed composition would only be greater, making our current analysis conservative. Thus replication-associated mutation cannot account for either the leading strand excess of A over T that is observed either in *S. aureus* intergenic regions (Table 1) or along the entire chromosome (Figure 3).

3) Evidence for skew being independent of replicatory strand. First and second codon sites, where the greatest AT skew is found, have signs of skew corresponding to the direction in which they are transcribed, irrespective of which replicating strand they lie on (Figure 5). This suggests replication-induced mutation is not contributing to the observed AT skews since such mutation would be expected to oppositely impact the magnitude of skew in the leading and lagging segments of a single DNA molecule.

These results indicate that the strong AT skew seen on the leading strand is dominantly owing to a strong strand bias, with the great majority of genes co-oriented with the replicatory fork (the leading strand contains 78% of the core coding content in *S. aureus*) and an abundance of A over T in coding sequence.

Avoidance of stop codons leads to AT skew at first codon positions

Having established that mutation is not the primary cause of the observed AT strand bias, we sought to determine what selective forces might be responsible. The principle challenge appears to be to explain why AT skew is so profound at first sites in codons, even when compared with codon second sites. As stop codons start with T and cannot feature within the coding sequence, their avoidance provides a potential component of the unusual first site AT skew.

While we are able to make an *a priori* assumption regarding stop codon usage (since, by definition, they cannot be included within the body of a gene), we have no prior expectation concerning amino acid usage. It was therefore necessary to measure the AT skew resulting from biased gene distribution in *S. aureus* in the absence of selection on amino acid-encoding codons. To this end we simulated coding sequences preserving the discrepancy in coding content between the two replicating strands of *S. aureus*. For each simulation, the same number of codons as seen in a given replicatory strand were reconstructed based on the intergenic nucleotide frequencies within that strand, but with the caveat that stop codons were not permitted. Intergenic base frequencies were used to derive codons in order to control for any effects of the baseline nucleotide content as well as any mutational contribution to coding content within the chromosome. AT skew was then calculated in first and second sites for each of the 10,000 randomized coding sequences. These simulations quantified the AT skew expected to result purely from the avoidance of stop codons, indicating that randomized coding sequences display significant AT skew in first positions (Table 3). Thus, we would expect a lack of stop codons to contribute significant AT skew in first positions given such a discrepancy in coding content between the two replicating strands, even with a complete lack of selection on amino acid content. However, the magnitude of the effect owing to stop codon avoidance is unable to explain the full magnitude of the skew that we observe.

Selective pressure for cost-effective amino acids leads to AT skew

To explain the residual AT skew at first sites and all of the AT skew at second sites left unexplained by the avoidance of stop codons in reading frames, we investigated the possibility of further selection within coding sequences to decrease T. The mean codon usage in the randomized coding sequences (where codons are drawn randomly in proportion to the intergenic nucleotide frequencies within the same replicating strand) represents a null expectation of codon usage in the absence of any selection on amino acid content. Comparing this null with the observed codon usage in the TW20 chromosome allows for direct quantification of the over- or under-usage of a given codon (z, see Methods). A positive z-score for a given amino acid indicates that amino acid is more commonly used within the *S. aureus* genome than would be expected according to our null model of codon usage, while a negative z indicates that amino acid is under-used. Such an approach reveals that T-rich codons are in fact highly under-represented in the *S. aureus* chromosome, with an enhanced avoidance of T in the gene-rich leading strand (Figure 6). What selective force could account for such a paucity of T in first and, to a lesser extent, second sites? We hypothesize that it might reflect unusual features of the amino acids that start with T.

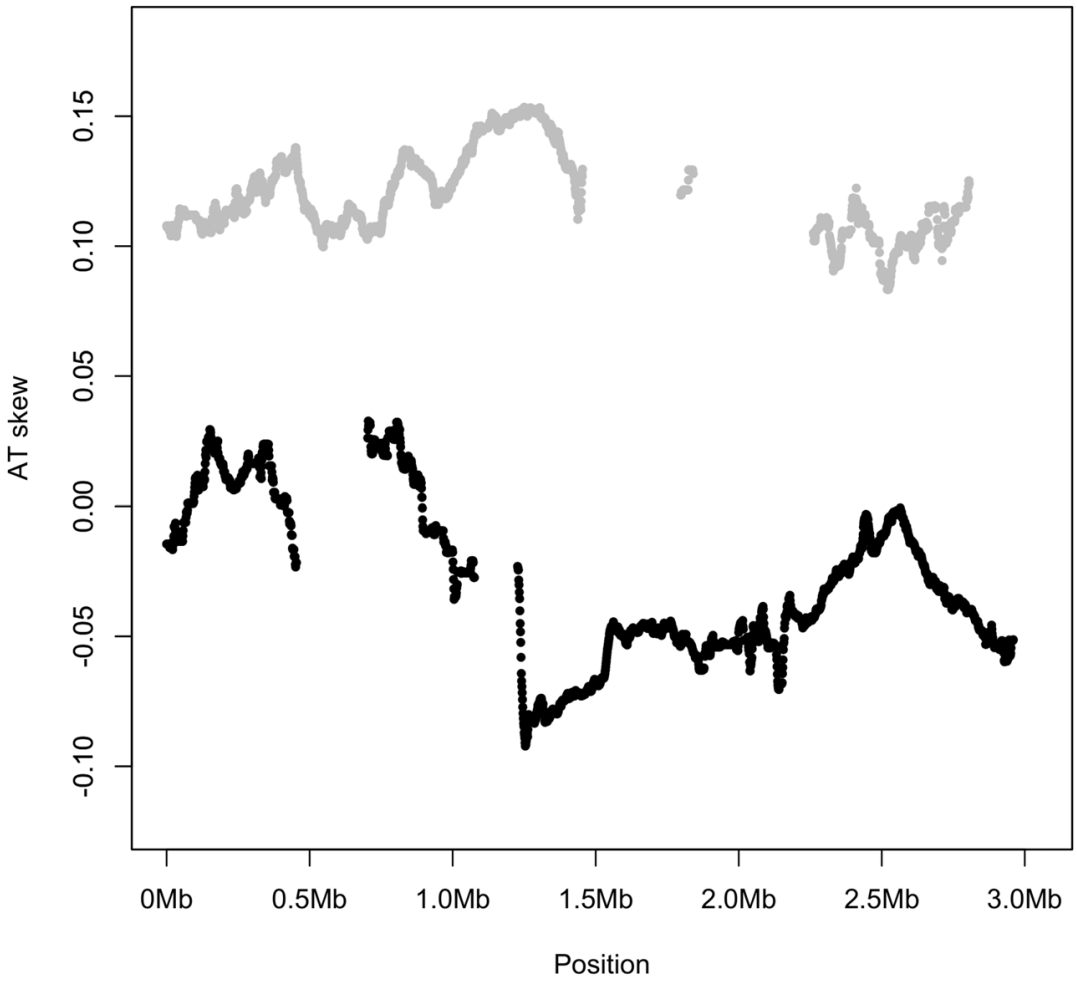

Figure 5. AT skew is similar along the entire length of genome for all noncomplementary genes (gray), while all complementary genes (black) display a similar AT skew. Genic AT skew was calculated with respect to the published strand, using overlapping windows of 300 kb and 1 kb steps where a data point was plotted if the number of bases in protein-coding genes in a window reached at least 30,000. Thus the skew shown for the noncomplementary genes (gray) corresponds to their coding sense direction whereas the skew shown for complementary genes (black) corresponds to their coding antisense direction. Skews were plotted this way to visually distinguish between different segments of the replicatory strands. If AT skew were primarily induced by replication, leading strand genes (the first half of the gray strand and the second half of the black strand) should show similar skews, and lagging strand genes should show roughly uniform skews opposite in sign to the leading strand genic skews. However this is not the case and *S. aureus* genes show AT skew corresponding to the direction in which they are transcribed, or the direction in which their amino acids are encoded.
doi:10.1371/journal.pgen.1002283.g005

Table 3. Simulated AT skew (that which can be accounted for by avoidance of stops in coding frames) in first and second codon positions contrasted with skews observed in *S. aureus* among all protein-coding genes.

Strand	Sites	Observed AT skew (TW20)	Mean simulated AT skew	P
Leading	1st positions	0.2403	0.1767±0.0016	<0.0001
	2nd positions	0.0717	−0.0731±0.0016	<0.0001
Lagging	1st positions	0.2081	0.0850±0.0030	<0.0001
	2nd positions	0.0245	−0.1250±0.0029	<0.0001

Both simulated and observed skews are given in the sense direction.
doi:10.1371/journal.pgen.1002283.t003

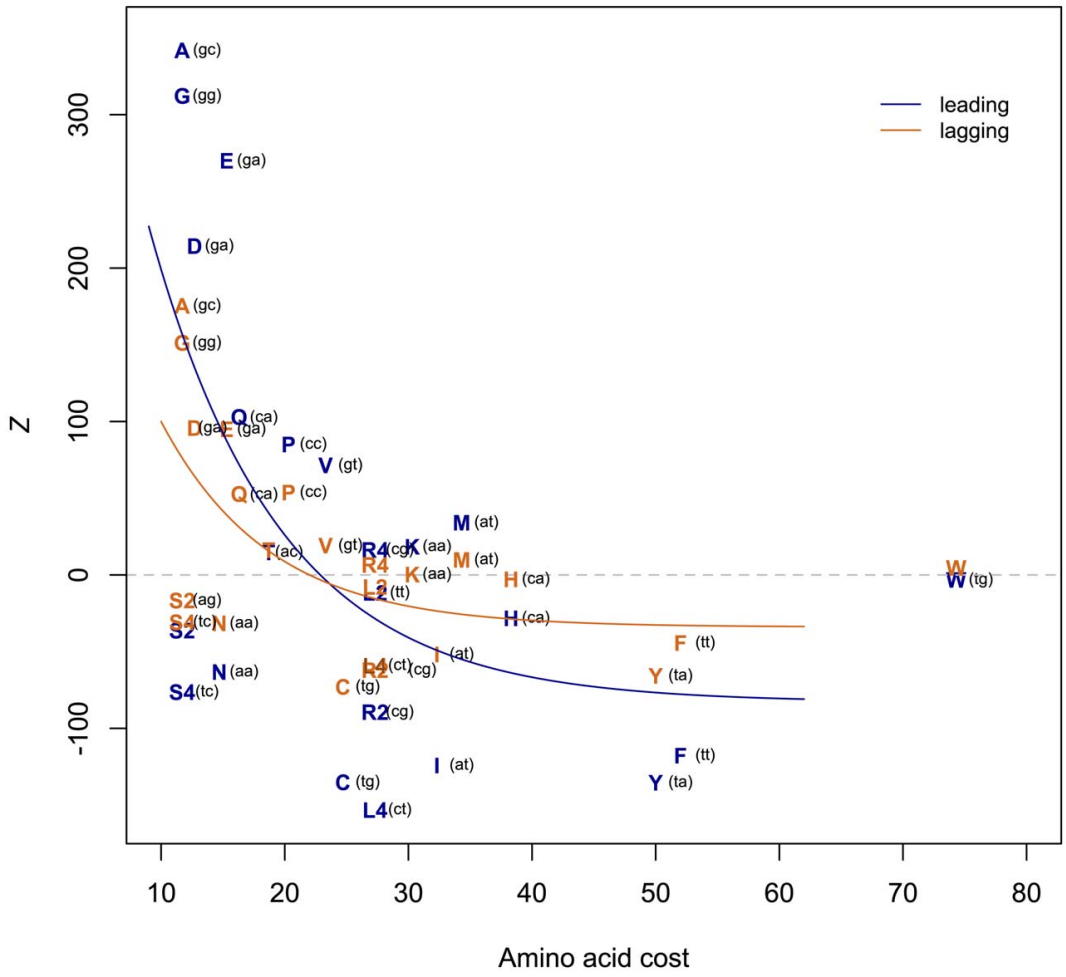

Figure 6. T-rich codons are under-represented in *S. aureus*. A positive *Z* represents over-usage, a negative *Z* under-usage. The negative correlation between *Z* and amino acid cost implies avoidance of costly (T-rich) amino acids, with the effect stronger on the leading strand (leading Spearman's rho-0.466, one-sided P = 0.013; lagging rho -0.468, P = 0.012). The first two codon positions of each amino acid are indicated in parentheses. Amino acid costs from{Akashi, 2002 #65} Akashi and Gojobori 2002 [29].
doi:10.1371/journal.pgen.1002283.g006

T-starting amino acids mostly belong to the shikimate pathway and are thought to have been those most recently added to the genetic code [28]. As late-added amino acids tend to be expensive to manufacture [29] and shikimates in particular tend to have complex chemical structures, might it be that T avoidance reflects nothing more than selection against the use of costly amino acids? We find that there is a significant negative correlation between $\mathcal{Z}$ and amino acid cost as given in Akashi and Gojobori 2002 [30] (Figure 6), indicating that expensive amino acids are indeed under-used, thus accounting for some of the paucity of T. We have repeated this analysis using alternative cost measures [31] and obtained similar results for six out of eight cost schemas (Figures S2, S3, S4, S5, S6, S7, S8, and Table S2). The two measures ($R_{glucose}$ and molecular weight) that do not provide significant correlations between $\mathcal{Z}$ and amino acid cost are perhaps expected to be less revealing: $R_{glucose}$ correlates neither with previous cost measures nor with amino acid substitution rate, while molecular weight does not take into account metabolic networks relating to amino acid production [31].

High gene strand bias can also account for atypical AT skew in other Firmicutes

We have shown that a lack of stop codons and avoidance of costly amino acids in asymmetrically distributed open reading frames can in large part account for the positive AT skew in *S. aureus*. Could similar mechanisms produce the unusual AT skews seen across other Firmicutes? Using phylogenetically independent contrasts (see Methods), we note that, among Firmicutes, an increase in the degree of gene strandedness from one species to another also results in a proportional increase in the extent of positive AT skew (Figure 7). It is therefore likely that the high degree of gene strand bias similarly explains the atypical patterns of AT skew in other Firmicutes.

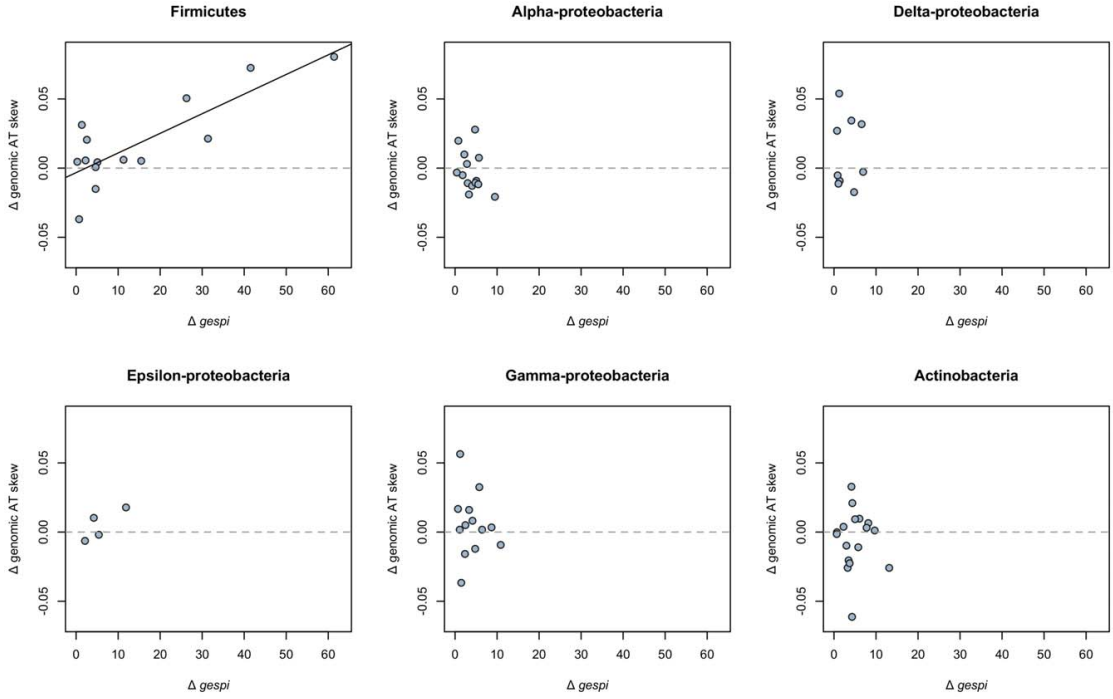

Figure 7. Gene strand bias predicts the extent of positive AT skew in the Firmicutes. Among Firmicutes, an increase in the degree of strand bias (positive Δ*gespi*) between terminal node species results in an increase in genomic AT skew, measured with respect to the leading strand (regression slope $P<0.001$, $r^2=0.65$; a one-sided binomial test for association between positive Δ*gespi* and positive Δgenomic AT skew is also significant at $P<0.01$). Consistent with our model that strand bias dictates the extent of positive AT skew, and therefore no change in *gespi* should not result in a change in AT skew, we cannot reject that the y-intercept goes through 0 (intercept $P=0.628$). There is no detectable significant relationship between changes in strand bias and genomic AT skew among members the following phyla (regression slope P values given in parentheses): the Alpha-proteobacteria ($P=0.196$), Delta-proteobacteria ($P=0.984$), Epsilon-proteobacteria ($P=0.185$), Gamma-proteobacteria ($P=0.654$), and Actinobacteria ($P=0.87$).
doi:10.1371/journal.pgen.1002283.g007

As a contrast to the Firmicutes, we performed a similar analysis on phylogenies (Tables S3, S4, S5, S6, S7, S8) of the Gram negative Alpha-proteobacteria [32], Delta-proteobacteria [33], Epsilon-proteobacteria [34], Gamma-proteobacteria [35], as well as the Gram positive Actinobacteria [36]. Although the species sampled from these phyla tend to display typical (negative) genomic AT skews, it is possible that the degree of strand bias within these genomes nevertheless modulates the magnitude of these AT skews due to either avoidance of stop codons or selection on amino acid cost. The additional lineages do not however reveal any pattern of regression of ΔATskew on Δ*gespi*, the latter indicating changes in gene strand bias (Figure 7). This is not unexpected since lack of large values of strand bias between terminal node pairs of the non-Firmicute phyla (resulting in lack of large differences in strand bias) means that the points all scatter around 0.

Discussion

Recently an interest has emerged in whether certain sites in bacterial chromosomes commonly thought to be nearly neutral are in fact under selection as regards their nucleotide content. Two studies [23,24] both used SNP profiles to estimate the GC content in possibly neutral sites at mutational equilibrium and showed the observed GC:AT bias greatly differs from that expected under the influence of mutation alone, consistent with previous reports of mutational pressure towards AT in *E. coli* [37]. What these studies were unable to explain, however, was what selective forces might be biasing nucleotide content at third codon and intergenic sites, leaving open the possibility that biased gene conversion and not selection might be acting. Here we also investigate a feature of bacterial chromosomes commonly presumed to be mutational, AT skew, and test whether mutation or selection is responsible for generating the unusual AT skews in *S. aureus*. Not only do we show that the atypical AT skew pattern in *S. aureus* is not due to mutational bias, but we are able to delineate to some degree what mode of selection is occurring (at least in terms of coding sequence), and to what end, in order to explain the observed skew pattern.

We find the mutational effect on AT skew in *S. aureus* (and in another Firmicute, *B. anthracis*), as derived from intergenic SNPs some distance away from coding sequence, to be inconsistent with, and poorly explanatory of, the observed base composition. Fourfold degenerate sites and intergenic regions display little skew, and intergenic SNP profiles do not support a replication-induced mutational origin of AT skew. In addition to fourfold sites, intra-operonic intergenic regions also display very weak AT skews, and hence any transcriptional effect is likely to be weak. Instead our results support a selectionist basis for compositional bias in *S. aureus* in which AT skew, the majority of which is observable at

first and second positions in the sense direction, results from selection at both the translational level and on gene position. The avoidance of stop codons and codons encoding costly amino acids accounts for a substantial proportion of the skew in first and second codon positions because the majority of genes are on the leading strand. However, we are unable to accurately quantify the contribution of the avoidance of costly amino-acids, because the cost estimates used [31] are only approximate. Nevertheless, we can describe a relationship between the intensity of selection against costly amino acids and the magnitude of skews (Figures S9, S10, and Table S9). Codons encoding more costly amino acids tend to be AT-rich [29] and we observe that, on average, AT-rich genomes encode more costly amino acids (Figure S9A). However, the average cost of amino acids in AT rich genomes while high, is not as high as expected given the AT pressure (Figure S9B). This we interpret as evidence for more efficient selection against costly amino acids in GC-poor strains, which in turn contributes to a higher AT skew (Figure S10).

We further show a phylogenetically controlled positive association between the extent of gene strand bias and positive genomic AT skew across the Firmicutes, indicating that strand bias is likely responsible in part for the atypical AT skews seen across this phylum. Our failure to detect such a relationship in non-Firmicute phyla may in part be due to a lack of genomes in these phyla with very high strand bias, leaving only increases in strand bias of smaller magnitude to investigate and thus much noisier data sets (Figure 7). The pattern of $\Delta gespi$ versus ΔATskew observed for the Firmicutes is similar to the patterns observed in other phyla when considering the region $0>x<10$ (Figure 7), meaning large differences in strand bias between terminal node pairs are required to be able to detect a relationship between the two quantities. This may be why we only see an effect in the Firmicutes, where strand bias is high enough to leave a clear impact upon the magnitudes of genomic AT skews. We conclude that if there is a relationship between *gespi* and AT skew in non-Firmicutes, our method is not sensitive enough to detect it.

Our results leave several mysteries. First, why do species differ in the degree of strand bias and why is it so high in many Firmicutes? These issues remain enigmatic. A simple model supposes that in fast replicating species the chance of DNA and RNA polymerases colliding must be higher than in slow replicating species. There is, however, no correlation between growth rate and gene strand bias [38]. Rather the higher biases are typically found in chromosomes containing two different (possibly strand-dedicated) DNAP α-subunits at the replication fork which may render them more vulnerable to polymerase collisions [15]. It has also been suggested that strand bias reflects gene essentiality rather than the level of expression [39] although, again, this does not explain the unusually high level of strand bias in Firmicute chromosomes.

Further, while both the observed AT skew in non-coding sites and the pattern of SNPs in intergenic sequence cannot explain the skew seen across the leading strands as a whole, the two approaches are also inconsistent with each other. The relative mutation rates calculated from intergenic SNPs indicate that mutation is acting to bias T over A in intergenic sites, which is the typical direction that AT skew takes in most (e.g. many non-Firmicute) bacteria, suggesting less variable skew-related mutational profiles among bacteria than is commonly assumed. However, intergenic sites on the leading strand have a weak bias in the opposite direction. Such a leading strand bias is consistent with leading strand coding sites also showing slightly higher bias than lagging strand coding sites (Table 1).

What could account for the discrepancy between mutational biases and observed base frequencies at putatively neutral sites? One possibility is that these sites are not yet at mutational equilibrium. This could occur if, for example, there were until recently some unannotated small protein coding genes in the "intergene" spacer. These new pseudogenes would take an appreciable time to reach mutational equilibrium and could well leave a trace of A>T skew if they tended to be on the leading strand. However, in this case it is curious that intergenic skew, intra-operonic skew and skew at four-fold degenerate sites all show a weak A>T bias. An alternative is that the weakly positive intergenic AT skews could reflect ongoing selection. One possibility is that there exists unannotated coding sequence, which, if enriched on the leading strand, would contribute a net A>T skew. Neither missing gene model can explain why skew at four fold degenerate sites is of the same magnitude as in putative intergene spacer. In addition, if mutation alone dictated intergenic AT skews, leading intergenic spacers should skew to roughly -0.4 (Table 2) according to our estimates of mutational equilibria. Given that the average leading AT skew in *S. aureus* coding sequence is approximately 0.1 across all three codon positions (Table 1), the vast majority of intergenic spacers would need to be unannotated protein-coding sequence in order for missing genes to be able to explain the observed leading intergenic AT skew of 0.0276 (Table 1), a highly untenable scenario. Removal of the few regions with outlier AT skew values does not substantially impact the intergenic AT skew (Figure S11), suggesting that even if we are missing some genes their contribution to skew cannot explain the overall bias. What exactly is generating weakly positive AT skews in leading intergenic regions remains a mystery, but this analysis adds weight to the growing evidence [see e.g. 19,23,24,37,40] that "neutral" sites in bacterial chromosomes may not be quite so neutral after all.

Methods

Sequence

The complete annotated genome of *S. aureus* subsp. *aureus* TW20, accession number FN433596 [41] was downloaded from the EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/). Data analysis was carried out using Tcl and Perl scripts and statistical analyses were done in R 2.9.0 [42]. Core and non-core regions were delineated as in Harris *et al.* 2010 [17]. Coding sequences labeled as gene remnants or pseudogenes were excluded and both known and putative genes were considered. As we considered intergenic regions to be indicative of mutational pressures, all intergenic regions were subject to a length restriction of 500 bp to decrease the possibility of unannotated genes, and 60 bp were trimmed from each end of all intergenic regions, as these regions display distinct AT skew patterns deviating from that induced by replication alone (see Results and Figure 4). This is most likely explained by ATG initiation context definition (at the 5′ end) and termination sequences (at the 3′ end).

Defining operons

Distinguishing which intergenic regions lie within operons should help reveal the contribution of transcriptionally-induced mutation to AT skew as such regions should be more likely to be transcribed. The operon structure of *S. aureus* strain MSSA476 [43] was used to deduce operons within strain TW20. Operonic protein-coding and RNA genes in MSSA476 were extracted from NCBI RefSeq NC_002953 (http://www.ncbi.nlm.nih.gov) and were matched via BLAST 2.2.24 against all TW20 protein sequences and gene-encoded RNA sequences respectively. Orthologous operonic genes in TW20 were taken to be those with at least 90 percent identity and an e-value of less than 0.0001. Due to

the close evolutionary relationship between the two strains most matches were unambiguous. Matches to pseudogenes and genes in the TW20 non-core chromosome were excluded. TW20 operons were then deduced from these orthologs with the additional caveat that genes within a putative operon be adjacent and transcribed in the same direction. In several cases a gene not contained in an MSSA476 operon was detected inserted into the intra-operonic TW20 intergenic sequence but on the strand opposite to the operonic genes. In these cases the operon was included.

SNP analysis

To determine whether the observed AT skew deviates from that expected under mutation alone, we require an estimation of nucleotide content, and the resulting AT skew, at mutational equilibrium. 140 singleton SNPs from core ex-operonic intergenic sites at least 60 bp away from a gene boundary were isolated from what is currently the largest SNP dataset for any bacterial species, that comprising 63 *S. aureus* ST239 isolates [17]. In order to check for the possibility that some of these SNPs were called erroneously, we went back to the original read data for each of these 140 SNPs individually. These data revealed an average of 20-fold coverage, a maximum of 46-fold coverage, and a minimum (in 4 SNPs) of 12-fold coverage (Table S10). Furthermore, in 126/140 (90%) of cases, the assigned SNP was consistent in all mapped reads. Of the 14 remaining SNPs, a single inconsistent read was noted in 13 cases, and two inconsistent reads noted in one case. Given a sequencing error rate of 0.5% per sequencing reaction, the maximum probability that any SNP has been assigned by error (that is called consistently in at least 12 reads) is of the order of $[(0.005\times0.3')\hat{}\ 11)] \approx 2\times10^{-31}$. Thus analysis of singleton mutations is an excellent indication of new mutations and does not reflect sequencing errors (see also Results).

These SNPs were used to estimate the mutational profile of *S. aureus* in ex-operonic intergenic sites. Singletons are SNPs which are seen only once throughout all sequenced isolates. Such SNPs are more likely to represent recent mutational events which selection has not yet had time to act upon, and thus only singletons were considered in order to orientate the direction of changes and minimize the possibility of selection or multiple hits. As all other lineages (aside from that with the singleton) have the same nucleotide at the given location, the assignment of the ancestral state is unambiguous.

As we find the sample size of 13 SNPs falling within intra-operonic intergenic regions too small to calculate the relative mutation matrix for intra-operonic sites, we considered SNPs in ex-operonic intergenic sites only. On both strands singletons were isolated in intergenic sites outside operons and relative rates of mutation were calculated for the leading and lagging strands separately. Nucleotide frequencies at mutational compositional equilibrium were derived from the relative mutation rates by considering that at compositional equilibrium, the loss of any given nucleotide must equal the net gain of that nucleotide at other sites:

$$f(A)r_{AT}+f(A)r_{AC}+f(A)r_{AG}=f(T)r_{TA}+f(C)r_{CA}+f(G)r_{GA}$$
$$f(T)r_{TA}+f(A)r_{TC}+f(A)r_{TG}=f(T)r_{AT}+f(C)r_{CT}+f(G)r_{GT}$$
$$f(C)r_{CT}+f(A)r_{CA}+f(A)r_{CG}=f(T)r_{TC}+f(C)r_{AC}+f(G)r_{GC}$$
$$f(G)r_{GT}+f(A)r_{GC}+f(A)r_{GA}=f(T)r_{TG}+f(C)r_{CG}+f(G)r_{AG}$$

where f(i) is the frequency of site i and r_{ij} is the rate of change from i to j per site i as measured in the extant sequence. The above equilibrium equations were solved simultaneously using Maxima 5.21.1 [44] to yield equilibrium nucleotide frequencies. These equilibrium frequencies were used to calculate the AT skew in ex-operonic intergenic sites expected to result purely from replicational mutation at compositional equilibrium.

Similar mutational equilibrium analyses were performed on polymorphism data from *B. anthracis* and *S. typhi*. Intergenic singleton SNPs were extracted from alignments of 18 fully and partially sequenced *B. anthracis* strains [23] and from the intra-haplotype or haplotype-specific age groups for *S. typhi* SNP data [45]. For both organisms, only intergenic regions under 500 bp were considered and singletons were only called when sequence data was available for all strains and the SNP at least 60 bp away from a gene. Observed intergenic nucleotide content and AT skew were calculated using NCBI RefSeqs NC_003997 and NC_003198.

To obtain an approximate measure of the robustness the sign of the equilibrium AT skew indicated by the singleton SNP populations, the intergenic (ex-operonic in the case of *S. aureus*) SNPs were bootstrapped. For each species, the intergenic SNPs used to compute the mutational equilibrium were resampled with replacement 1000 times, and the equilibrium state recalculated as above after each resampling, to yield 95% bootstrap intervals for the equilibrium AT skew estimate.

Randomizations

Selection against stop codons within asymmetrically distributed genes could necessarily impose some amount of AT skew as T might be underrepresented relative to A within first codon sites. We wished to measure the AT skew which results in *S. aureus* from selection on gene position alone while preventing any selection on amino acid content, which might further increase or decrease the amount of T relative to A within genes in *S. aureus*, from biasing this measurement. Randomized coding sequences provide a means of estimating the AT skew that would result from the biased gene orientation seen in *S. aureus* even under a complete lack of selection for amino-acid usage. As both GC content and replication-associated mutational biases can modulate the amino acid content of proteins [46,47], the baseline nucleotide frequencies of the leading and lagging strands of the TW20 chromosome could favor the presence of certain codons while disfavoring others. A null was devised in which nucleotides were sampled in proportion to their frequency in intergenic regions, which should be neutral or weakly selected, thus controlling for the baseline nucleotide content of the genome as well as any mutational effect on skew. 10,000 protein-coding sequences containing the same number of amino acid-encoding codons as in the leading and lagging strands of the TW20 chromosome were simulated using codons derived from the intergenic nucleotide frequencies in the relevant strand of the TW20 chromosome. Stop and start codons were excluded from randomized sequences. Amino acids with six codons were considered as two separate amino acids—a 4-block and a 2-block—since the frequency of individual nucleotides could differentially influence the usage of these two codon blocks. The resulting AT skew in randomized chromosomes was calculated in first and second sites as (A−T)/(A+T) with respect to the sense direction.

Selective patterns of amino acid usage may, depending on the nucleotide frequencies of the codons involved, also shape AT skew. Determination of whether individual amino acids are over- or under-used in relation to the above null is reflected in the Z score for each amino acid (aa):

$$Z_{aa}=\frac{[\text{Observed}-\text{expected usage of aa}]}{SD_{aa}}$$

where the expected usage is the mean usage of that amino acid amongst the 10000 simulated coding regions, the observed usage is that seen in the TW20 chromosome and the standard deviation is that observed through the randomizations. This normalizes for variance seen due to amino acids occupying differing amounts of codon space, controls for the effect that genomic GC content may have on individual codon usage, and allows for comparison of over- or under-usage across different amino acids.

Gene strandedness and genomic AT skew across Firmicutes

If gene strand bias is responsible for positive AT skews not just within *S. aureus* but across the Firmicutes, we expect a positive association between gene strandedness and genomic AT skews across the phylum. A simple test for correlation between these two quantities across a wide sampling of Firmicute species might, however, falsely infer a relationship between the two due to over-representation of sequence information in closely related genomes. We therefore investigated the relationship between strand bias and genomic AT skew using phylogenetically independent contrasts. Differences in strand bias and leading genomic AT skew were calculated for phylogenetically independent pairs of terminal node species in a phylogeny of Firmicutes [48] (Table S3) with the expectation that if strand bias does dictate the extent of positive AT skew, an increase in strand bias between species should also result in an increase in AT skew. *Gespi* values, calculated according to de Carvalho & Ferreira 2007 [49], were used as indicators of the degree of strand bias among these species, with a higher *gespi* indicating a greater degree of strandedness. As a counterpoint to the Firmicutes, similar analyses were performed on phylogenies of the Gram negative Alpha-proteobacteria [32], Delta-proteobacteria [33], Epsilon-proteobacteria [34], Gamma-proteobacteria [35], and a phylogeny of the Gram positive Actinobacteria [36] (Tables S4, S5, S6, S7, S8).

Supporting Information

Figure S1 *B. anthracis* and *S. typhi* both show fluctuations in AT skew in intergenic regions at gene boundaries. AT skew at each position was calculated from the nucleotide content measured across all intergenic regions at that position relative to the gene start or end as appropriate. All intergenic regions were considered in the direction of transcription of the relevant gene.
(DOC)

Figure S2 Z versus amino acid cost using alternative cost measure $A_{glucose.}$. A positive $\mathcal{Z}$ represents over-usage, a negative $\mathcal{Z}$ under-usage. Correlation between Z and amino acid cost, Spearman's rho: leading strand, -0.376, one-sided P = 0.038, lagging strand rho, -0.399, P = 0.031.
(DOC)

Figure S3 Z versus amino acid cost using alternative cost measure $R_{glucose.}$. A positive $\mathcal{Z}$ represents over-usage, a negative $\mathcal{Z}$ under-usage. Correlation between Z and amino acid cost, Spearman's rho: leading strand, one-sided P = 0.553, lagging strand, P = 0.553.
(DOC)

Figure S4 Z versus amino acid cost using the alternative cost measure of Craig and Weber energy. A positive $\mathcal{Z}$ represents over-usage, a negative $\mathcal{Z}$ under-usage. Correlation between Z and amino acid cost, Spearman's rho: leading strand, -0.578, one-sided P = 0.002, lagging strand rho, -0.566, P = 0.002.
(DOC)

Figure S5 Z versus amino acid cost using the alternative cost measure of Craig and Weber steps. A positive $\mathcal{Z}$ represents over-usage, a negative $\mathcal{Z}$ under-usage. Correlation between Z and amino acid cost, Spearman's rho, leading strand, -0.450, P = 0.016, lagging strand rho, -0.484, P = 0.009.
(DOC)

Figure S6 Z versus amino acid cost using the alternative cost measure of Wagner fermentative costs. A positive $\mathcal{Z}$ represents over-usage, a negative $\mathcal{Z}$ under-usage. Correlation between Z and amino acid cost, Spearman's rho, leading strand, -0.373, P = 0.040, lagging strand rho, -0.411, P = 0.026.
(DOC)

Figure S7 Z versus amino acid cost using the alternative cost measure of Wagner respiratory costs. A positive $\mathcal{Z}$ represents over-usage, a negative $\mathcal{Z}$ under-usage. Correlation between Z and amino acid cost, Spearman's rho, leading strand, -0.584, P = 0.002, lagging strand rho, -0.548, P = 0.003.
(DOC)

Figure S8 Z versus amino acid cost using alternative cost measure of molecular weight. A positive $\mathcal{Z}$ represents over-usage, a negative $\mathcal{Z}$ under-usage. Correlation between Z and amino acid cost, Spearman's rho, leading strand, P = 0.119, lagging strand, P = 0.079.
(DOC)

Figure S9 The observed cost of amino acids encoded in GC-poor genomes is lower than expected, suggesting more efficient cost selection in AT-rich bacteria. Coding sequences were simulated taking into account the GC contents of individual codon positions in the given strain and stop codon avoidance, with the number and length of simulated sequences based on observed values, assuming no GC or AT skew. (A) Box plot for 105 genomes, light blue: median simulated, dark blue: median observed Akashi and Gojobori [29] biosynthetic cost of amino acids encoded in genes, the lower and upper quartiles are shown in gray. (B) The factor by which the mean observed amino acid cost values are lower than expected correlates with genomic GC content. Orange: Firmicutes, Spearman's rho = -0.840, P$<2.2\times10^{-16}$, black: non-Firmicutes, Spearman's rho = -0.768, P$<2.2\times10^{-16}$, Firmicutes and non-Firmicutes together: Spearman's rho = -0.871, P-value $<2.2\times10^{-16}$.
(DOC)

Figure S10 AT skews correlate with the intensity of selection against costly amino acids. The ordinate shows the mean of AT skews calculated for individual protein coding genes in their sense direction in a given genome, for which the cost ratio was calculated as in Figure S9. Orange: Firmicutes, black: non-Firmicutes; see Table S2 for statistical data.
(DOC)

Figure S11 Outlier AT skew values are not responsible for the positive AT skews seen in ex-operonic intergenic regions. For each AT content observed in such an intergenic region in the TW20 genome, 1000 randomized sequences were created by shuffling the total nucleotide content in ex-operonic intergenic sequences 1000 times, and each time the shuffled sequence was repartitioned into intergenic regions containing the same AT contents as in the observed genome. The 95% confidence interval (black points) was calculated from these simulated sequences to determine which observed ex-operonic intergenic AT skew values (green points) were outliers (green filled points falling outside the 95% confidence interval). The leading AT skew in ex-operonic intergenic

sequences with outliers removed (0.0257) is very similar to the same calculation inclusive of outliers (0.0276).
(DOC)

Table S1 a. Relative mutation rates of nucleotide i to j per site i for intergenic sites were calculated from singleton SNPs for *B. anthracis* (gray rows) and *S. typhi* (white rows). All rates are shown with respect to the leading strand and derived from the following leading strand SNP counts, where XY indicates a change from nucleotide XY: *B. anthracis* SNPs, AG 9 GA 11 GC 1 CG 1 GT 4 TA 2 TC 20 TG 2 CA 3 AC 3 AT 5 CT 15. *S. typhi* SNPs, AG 6 GA 15 GC 0 CG 0 GT 3 TA 0 TC 4 TG 0 AC 0 CA 1 AT 0 CT 14. b. Current observed intergenic AT skew contrasted with SNP-derived intergenic equilibrium AT skews for *B. anthracis* and *S. typhi*. All skews are given with respect to the leading strand. 95% bootstrap intervals are shown in parentheses. That *B. anthracis* does not display a consistently negative bootstrap interval is a consequence of at least two factors. Firstly, the sample size (76 SNPs) used to derived the mutational equilibrium is small compared to that used for *S. aureus* (140 SNPs). Secondly, the alignments used to derive the *B. anthracis* SNPs come from several independent sequencing efforts and we are unable to verify the sequence qualities. As for *S. typhi*, the even smaller sample size of 43 SNPs leaves many mutational categories unrepresented and leads to inflated bootstrap intervals.
(DOC)

Table S2 Spearman rank correlations between z and amino acid cost using alternative cost measures.
(DOC)

Table S3 Terminal node comparisons taken from a phylogeny of Firmicutes [48] used to calculate the difference in *gespi* and leading strand genomic AT skew (where more than one species is listed in a field, the average of those genomes was taken).
(DOC)

Table S4 Terminal node comparisons taken from a phylogeny of Actinobacteria [36] used to calculate the difference in *gespi* and leading strand genomic AT skew.
(DOC)

Table S5 Terminal node comparisons taken from a phylogeny of Alpha-proteobacteria [32] used to calculate the difference in *gespi* and leading strand genomic AT skew.
(DOC)

Table S6 Terminal node comparisons taken from a phylogeny of Delta-proteobacteria [33] used to calculate the difference in *gespi* and leading strand genomic AT skew.
(DOC)

Table S7 Terminal node comparisons taken from a phylogeny of Epsilon-proteobacteria [34] used to calculate the difference in *gespi* and leading strand genomic AT skew.
(DOC)

Table S8 Terminal node comparisons taken from a phylogeny of Gamma-proteobacteria [35] used to calculate the difference in *gespi* and leading strand genomic AT skew (where more than one species is listed in a field, the average of those genomes was taken).
(DOC)

Table S9 Relationships between mean CDS AT skews and cost ratios calculated for eight different amino acid cost measures. C&W: Craig and Weber, Wagner ferm.: Wagner fermentative costs, Wagner resp.: Wagner respiratory costs. Amino acid cost ratios were calculated as in Figure S9.
(DOC)

Table S10 Coverage data for the 140 singleton ex-operonic intergenic SNPs used in this analysis. 126/140 (90%) of the SNPs were consistent called in all mapped reads (marked with a *). The minimum of consistent mapped reads was 12 (in four SNPs), and there was one SNP with 12 consistent reads and one inconsistent. For the 14 SNPs with inconsistent reads, all showed only a single inconsistent read bar one (which had 2 inconsistent reads). Given a sequence error rate of 0.5%, a high level of coverage and high consistency between reads, the probability that any of these SNPs are errors is negligible.
(DOC)

Acknowledgments

We are grateful to Ruth Hershberg for supplying the *B. anthracis* alignments and Simon R. Harris for providing the original *S. aureus* read mapping data. We thank Matthew T. G. Holden for his consideration of the manuscript.

Author Contributions

Conceived and designed the experiments: CAC FH LDH EJF. Analyzed the data: CAC FH JMB LDH EJF. Contributed reagents/materials/analysis tools: CAC LDH EJF. Wrote the paper: CAC LDH EJF.

References

1. Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40: 318–325.
2. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13: 660–665.
3. Lobry JR, Sueoka N (2002) Asymmetric directional mutation pressures in bacteria. Genome Biol 3: RESEARCH0058.
4. Rocha EP, Danchin A, Viari A (1999) Universal replication biases in bacteria. Mol Microbiol 32: 11–16.
5. McLean MJ, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol 47: 691–696.
6. Frank AC, Lobry JR (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238: 65–77.
7. Mrázek J, Karlin S (1998) Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci U S A 95: 3720–3725.
8. Nikolaou C, Almirantis Y (2005) A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. Nucleic Acids Res 33: 6816–6822.
9. Tillier ER, Collins RA (2000) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol 50: 249–257.
10. Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. Trends Genet 13: 240–245.
11. Necşulea A, Lobry JR (2007) A new method for assessing the effect of replication on DNA base composition asymmetry. Mol Biol Evol 24: 2169–2179.
12. Rocha EPC, Touchon M, Feil EJ (2006) Similar compositional biases are caused by very different mutational effects. Genome Research 16: 1537–1547.
13. Morton RA, Morton BR (2007) Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes. BMC Genomics 8: 369.
14. Brewer BJ (1988) When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. Cell 53: 679–686.
15. Rocha E (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 10: 393–395.
16. Worning P, Jensen LJ, Hallin PF, Staerfeldt HH, Ussery DW (2006) Origin of replication in circular prokaryotic chromosomes. Environ Microbiol 8: 353–361.
17. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, et al. (2010) Evolution of MRSA During Hospital Transmission and Intercontinental Spread. Science 327: 469–474.
18. Rocha EP, Danchin A, Viari A (1999) Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Res 27: 3567–3576.
19. Molina N, van Nimwegen E (2008) Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res 18: 148–160.
20. Angov E, Brusilow WS (1994) Effects of deletions in the uncA-uncG intergenic regions on expression of uncG, the gene for the gamma subunit of the Escherichia coli F1Fo-ATPase. Biochim Biophys Acta 1183: 499–503.

21. Castillo-Ramírez S, Harris S, Holden M, He M, Parkhill J, et al. (2011) The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog 7: e1002129. doi:10.1371/journal.ppat.1002129.
22. Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, et al. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239: 226–235.
23. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001107. doi:10.1371/journal.pgen.1001107.
24. Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6: e1001107. doi:10.1371/journal.pgen.1001107.
25. Messer PW (2009) Measuring the Rates of Spontaneous Mutation From Deep and Large-Scale Polymorphism Data. Genetics 182: 1219–1232.
26. Liu X, Maxwell TJ, Boerwinkle E, Fu YX (2009) Inferring population mutation rate and sequencing error rate using the SNP frequency spectrum in a sample of DNA sequences. Mol Biol Evol 26: 1479–1490.
27. Achaz G (2008) Testing for neutrality in samples with sequencing errors. Genetics 179: 1409–1424.
28. Taylor FJR, Coates D (1989) The Code Within the Codons. Biosystems 22: 177–187.
29. Hurst LD, Feil EJ, Rocha EP (2006) Protein evolution: causes of trends in amino-acid gain and loss. Nature 442: E11–12. discussion E12.
30. Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proceedings Of the National Academy Of Sciences Of the United States Of America 99: 3695–3700.
31. Barton MD, Delneri D, Oliver SG, Rattray M, Bergman CM (2010) Evolutionary Systems Biology of Amino Acid Biosynthetic Cost in Yeast. PLoS ONE 5: e11935. doi:10.1371/journal.pone.0011935.
32. Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the alphaproteobacteria. J Bacteriol 189: 4578–4586.
33. Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, et al. (2010) Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. Syst Appl Microbiol 33: 291–299.
34. Takai K, Suzuki M, Nakagawa S, Miyazaki M, Suzuki Y, et al. (2006) Sulfurimonas paralvinellae sp. nov., a novel mesophilic, hydrogen- and sulfur-oxidizing chemolithoautotroph within the Epsilonproteobacteria isolated from a deep-sea hydrothermal vent polychaete nest, reclassification of Thiomicrospira denitrificans as Sulfurimonas denitrificans comb. nov. and emended description of the genus Sulfurimonas. Int J Syst Evol Microbiol 56: 1725–1733.
35. Williams KP, Gillespie JJ, Sobral BW, Nordberg EK, Snyder EE, et al. (2010) Phylogeny of gammaproteobacteria. J Bacteriol 192: 2305–2314.
36. Ludwig W, Euzéby J, Whitman WB (2011) Phylogenetic trees of the phylum Actinobacteria. In :Goodfellow M, Kämpfer P, Hans-Jürgen B, Trujillo ME, Suzuki K-i, et al. (2011) Bergey's Manual of Systematic Bacteriology. New York: Springer.
37. Balbi KJ, Rocha EPC, Feil EJ (2009) The Temporal Dynamics of Slightly Deleterious Mutations in Escherichia coli and Shigella spp. Molecular Biology and Evolution 26: 345–355.
38. Rocha EPC (2008) The Organization of the Bacterial Genome. Annual Review of Genetics 42: 211–233.
39. Rocha EP, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377–378.
40. Plague GR (2010) Intergenic transposable elements are not randomly distributed in bacteria. Genome Biol Evol 2: 584–590.
41. Holden MT, Lindsay JA, Corton C, Quail MA, Cockfield JD, et al. (2010) Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW). J Bacteriol 192: 888–892.
42. R Development Core Team (2005) R: A Language and Environment for Statistical Computing. ViennaAustria: R Foundation for Statistical Computing.
43. ten Broeke-Smits NJ, Pronk TE, Jongerius I, Bruning O, Wittink FR, et al. (2010) Operon structure of Staphylococcus aureus. Nucleic Acids Res 38: 3263–3274.
44. Maxima sourceforge.net (2009) Maxima, a Computer Algebra System. Version 5.21.1. http://maxima.sourceforge.net.
45. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, et al. (2008) High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nature Genetics 40: 987–993.
46. Lobry JR (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205: 309–316.
47. Mackiewicz P, Gierlik A, Kowalczuk M, Dudek MR, Cebrat S (1999) How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res 9: 409–416.
48. Wolf M, Muller T, Dandekar T, Pollack JD (2004) Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. Int J Syst Evol Microbiol 54: 871–875.
49. de Carvalho MO, Ferreira HB (2007) Quantitative determination of gene strand bias in prokaryotic genomes. Genomics 90: 733–740.

VI. Discussion

Interrogating whether a variety of encoded sequence features are under selection to enhance the efficiency of translation-related processes, I find that one type of selection for translational efficiency results from the need to balance energetic expenses. That is, the need for metabolically cheap to manufacture amino acids coupled with heavy loading of coding content onto the leading strand results in the anomalous AT skews in the Firmicutes. But is it just Firmicutes that are under selection for energetically cheap to manufacture coding sequences? The richer in AT content a genome is, the more metabolically expensive random coding sequences drawn out of the observed genomic nucleotide frequencies, with concomitantly greater deviation from expected costs observed in real genomes (Chapter V, Figure S9). We thus infer that there may be greater pressure in AT-rich genomes to optimize the metabolic efficiency of coding sequences.

Note that selection for metabolically cheap amino acids need not however result in genome-wide skews. We simply see such skews in the case of the Firmicutes as the strong gene strand bias causes asymmetric distribution of the A>T pressure in non-synonymous sites to the leading strand. But what are the reasons for this augmented gene strand bias? One possibility is that loading of genes onto the leading strand allows for replication fork maintenance. Both co-directional and head-on collisions between DNA and RNA polymerases have been shown to cause replication restart, although head-on collisions are more severe in that they also cause a significant slowing of replication (Merrikh et al. 2011). However, the need to grow fast cannot by itself explain loading of coding content on the leading strand, as no correlation between gene strand bias and growth rate is observed (Rocha 2008). Another possibility is that polymerase collisions might have an even greater detrimental fitness consequence if they compromise the production of essential genes, which might explain the preferential localization of such genes to the leading strand (Rocha and Danchin 2003). However, as all organisms have essential genes, this observation still leaves unanswered why gene orientation bias is stronger in some organisms (including Firmicutes) than others. Rather, what is perhaps the most comprehensive explanation for gene strand bias has to do with the physical components of the replication fork. It has been noted that while most bacteria use the *DnaE* gene product to synthesize both the leading and lagging strands, many low G+C Firmicutes (which also show high strand bias) synthesize only the lagging strand with DnaE and utilize PolC for replication of the leading strand (Rocha 2002). Thus the subunit encoded by *PolC* may introduce instability in the replication fork, and gene strand bias might enhanced to minimize head-on collisions if they are more deleterious to genome integrity than usual in PolC-utilizing organisms. Indeed, engineered inversions in *Bacillus subtilis* which alter the normal co-orientation of replication and translation lead to a stress response including activation of DNA repair mechanisms; nevertheless, disruption of genome

integrity and cell death are also observed (Srivatsan et al. 2010). Whether the fitness consequences of such inversions would be less deleterious in a non-Firmicute species remains to be seen.

The potentially deleterious role of collisions is also seen in the case of the ramp, where selection on ribosome-slowing positive charges at N-termini is postulated to serve as a kind of speed bump to prevent traffic jams between ribosomes (Tuller et al. 2011). The prevention of ribosomal traffic jams is hypothesized to enhance the efficiency of translation both locally, along the individual transcript, by preventing translation arrest (and the associated production of potentially toxic partial peptides) as well as globally by reducing the number of stalled ribosomes, freeing them up to translate other transcripts (Andersson and Kurland 1990; Gingold and Pilpel 2011). 5' traffic jams however have only been postulated to both exist, as well as to lead to ribosomal stalling. If there are only a few ribosomes per transcript at any point in time on most transcripts then the beneficial consequences of an elongatory 'ramp' would be hard to imagine, regardless of the mechanistic consequences of two ribosomes colliding. Further, that slowing of ribosomes followed by relative speeding will prevent traffic jams has also never been shown, only modelled (Mitarai et al. 2008; Tuller et al. 2010). This is important because it is unclear what the mechanistic consequences of collisions between two normally translating ribosomes actually are. Previous studies of ribosomal stalling during translation elongation have centred around stalling induced by other causes, such as conserved ribosomal stalling sequences (Sunohara et al. 2004) or tRNA depletion (Li et al. 2006), rather than by ribosomal collisions per se. Exactly how one ribosome responds to bumping into another ribosome, and the frequency with which this occurs on transcripts, are open questions. In addition, the fitness effects of such collisions might perhaps also be dependent on factors such as the rates at which the ribosomes are traveling and how densely they are packed. Indeed one study found that in bacteria, which undergo co-transcriptional translation, ribosomes which run into RNA polymerase can stimulate the latter's activity (Proshkin et al. 2010). Analogously, might one ribosome colliding into the next also promote translation rather than hinder it? Or perhaps ribosomal collisions result in minor slowing of translation, rather than its abortion? These types of mechanistic questions should ideally be addressed in any future hypotheses of translational traffic regulation.

Additionally, it has been suggested that ribosomal occlusion of transcript sequence can prevent mRNA degradation by altering stability and/or access to endonucleases (Deana and Belasco 2005). One study for example found greater amounts of protein produced from an enzyme loaded with naturally occurring rare codons at the 5' ends (and overexpressed from a strong plasmid) is consistent slow-travelling ribosomes preventing mRNA decay, thereby increasing the number of transcripts available to be translated compared to the coding sequence of a highly related isozyme whose most 5' codons are optimal (Kolmsee and Hengge 2011). Apart from

begging the question of whether spacing of ribosomes might be under selection to prevent mRNA degradation rather than collisions, this study also underscores a fundamental problem in interpreting experimental results. If the authors had not tested the action of endonucleases, the rise in protein production resulting in the slow construct may have been found to be instead consistent with a model where increased translational efficiency was a function of selection on rare codons to prevent ribosomal stalling, as per the ramp hypothesis (Mitarai et al. 2008; Tuller et al. 2010). Thus it remains difficult to unequivocally interpret the results of experiments where alteration of one feature, for example codon usage, in turn alters other linked properties such as RNA folding, decay rates, initiation rates (particularly due to transcript features at the 5' end), and so on. The myriad of possible side effects due to changes in these linked, sequence-derived features when making substitutions to coding sequences advises particular caution against over-interpreting the results of any single experiment. Drawing conclusions across experiments done in different species, moreover, may also be unwise if codon frequency and nucleotide content vary across species, as different binding energies and hence mRNA stabilities may emerge from the use of rare codons in different species.

The above discussion touches upon doubts concerning the mechanistic functioning of a translational, gene regulatory ramp. On a broader level, we can also question whether the ramp as conceived is an effective solution to the presumed problem at hand. It is not immediately apparent why an organism which can experience relatively strong levels of selection (e.g. yeast) could end up initiating translation more quickly than it is capable of undertaking. The ramp hypothesis presumes that such recurrent initiation results in a fitness defect severe enough that it must be selected against; however, curiously, in the proposed scheme the cell opts not for more intermittent translation initiation which might be made possible by a few changes in the transcript folding energy at 5' ends (e.g. Kudla et al. 2009), and/or possibly by decreasing the mRNA:protein ratio. Rather, according to the ramp hypothesis, the cell opts to select, across multiple functional levels of coding sequence (codons, folding, and amino acid charge) for supposed slowing features that will compensate for the maintained rates of initiation which are kept too rapid. If ribosomal jamming is indeed a problem impeding the translational efficiency of varied transcripts, the ramp seems an inefficient and roundabout solution.

As a final remark regarding the ramp, I reported that average positive charge use does not increase nearing N-termini in cytosolic (non-transmembrane) proteins. Yet it is still the case that positive charge is indeed used in a number of these cytoplasmic proteins, and a subset of which may in fact be front-loaded with more positive charges than are found on average in the cytosolic N-terminus of a membrane protein. In other words, it is theoretically possible that some non-membrane proteins might have 5, 10, 15 positive charges in their N-termini, so while there is no

increasing pattern when averaging across these globular proteins, this subset may even have more of a 'ramp' than any single random transmembrane protein. Whether we find evidence for such a ramp in yeast will need to be investigated, however it seems unlikely that such proteins will display a ramp in *E. coli* given that we detected no real ramp (i.e. 5' ribosomal excess) at all beyond a very short, initial excess likely more indicative of initiation than elongation. I would also note that the original study (Tuller et al. 2011) that claimed a role for positive charges in preventing ribosomal collisions did not consider such proteins independently as I have proposed. Rather, the evidence cited to support the claim that positive charges are selected to modulate ribosomal velocity was presented in the form of a correlation between average ribosomal density (calculated from all transcripts) and average positive charge use. It is exactly this average positive charge use which is shown in Chapter IV to be due to the orientation of membrane proteins rather than a gene regulatory ramp. Thus, it so far remains to be shown that any protein-level feature, such as positive charge, has been under selection for translational efficiency.

What about transcript-level features—are they under selection for efficiency of translation? The contribution of mRNA folding, on average, to slowing appears to be minimal, and I find no evidence that rare codons (i.e. both genomically rare and corresponding to rare tRNAs) cause significant slowing. Thus the possibility that such codons might be selected e.g. at the beginning of transcripts to regulate gene expression via modulation of the local translation rate (as proposed by e.g. Mitarai et al. 2008; Parmley and Huynen 2009; Clarke and Clark 2010; Tuller et al. 2010; Tuller et al. 2011) seems correspondingly unlikely. Rather, that aberrant codon usage is observed at 5' ends might be better explained in terms of selection on transcript structure which may be related to translation initiation (Eyre-Walker and Bulmer 1993; Kudla et al. 2009; Bentele et al. 2013). The analysis presented here does of course pertain to yeast cells, and it is possible different results regarding variation in codon speeds may be achieved in different organisms, particularly ones in which small population sizes may constrain the efficacy of selection. It is also possible that tRNA:codon proportionality varies throughout the cell cycle, thus possibly altering the supply of tRNAs to different codons and changing their elongation rates. Although models have been put forward to suggest that differential codon usage at varying points throughout the cell cycle is responsible for fluctuations in protein product levels (Frenkel-Morgenstern et al. 2012), it is still hard to envision that codon usage could limit the amount of expression of certain proteins if a) codons do not drastically differ in their elongation rates in vivo and b) initiation is normally rate-limiting in vivo (Laursen et al. 2005; Shah et al. 2013).

The result that codons do not differ markedly in their elongation rates is not necessarily inconsistent, however with a global efficiency hypothesis. A global efficiency argument need not require that different codons are translated at significantly different rates; it simply requires that

the codons used in highly expressed genes are provided in higher amounts so they do not become rate-limiting in protein production. In other words, selection on tRNA supply and codon demand might occur to allow cells to maximize their growth, especially at fast rates (Andersson and Kurland 1990). As Ikemura noted (1981), global optimization of codon usage is evolutionarily possible if a slow-growing organism comes under selection to grow faster, with tRNA and codon usage co-adapting in highly expressed genes. However, once at equilibrium, the fitness effect of any small mutation—such as a point mutation in a codon—is expected to be rather small. Thus cells may hover around a codon usage optimum, with the fitness effect of any particular synonymous substitution being negligible (overlooking intertwined factors such as mRNA folding) as long as the total cellular codon:tRNA proportionality is roughly maintained. Rather than selection to exacerbate rate differences between codons, selection for tRNA:codon proportionality in this case may act to ensure that no codon significantly limits elongation, particularly in highly expressed genes or at fast growth rates where the cell is presumably actively transcribing and translating at high levels. That codon speeds might be equalized via selection on their supply lines (tRNA levels) raises an important point which I wish to emphasize here: that codons do not significantly differ in elongation speeds one to the next does not necessarily mean lack of speed selection on codons.

Yet the claim presented in this thesis that codons do not differ in their translation speeds should not be over-interpreted. It would be ridiculous to argue that rates of different codons are $100.\overline{0}$ % equal. Such a scenario would require improbable levels of selection, especially given that stochasticity in supply and demand that may arise depending on what codons exactly are being translated at any given point in time. Some amount of variation in codon speeds may, perhaps, be even more likely given reports of population-level heterogeneity in microorganism gene expression profiles from one cell to the next (Taniguchi et al. 2010). These differences in gene expression relative to the population-constant tRNA supply could potentially give rise to slightly different translation speeds for a single given codon depending on which particular cell is being considered. Rather than asking if codon speeds differ, the question is perhaps better phrased as whether this magnitude of slowing difference from codon to codon is significant, not just statistically (in that they might account for a significant proportion of the total variation in ribosomal movement along transcripts) but, ultimately, biologically. The analyses presented here show that the contribution of codon usage to variations in translation speed along a transcript must be rather small compared to that of positive charge.

There are other limits to the current analysis. My investigations into the ribosomal slowing induced by sequence-level features also leave open the question of what the magnitude (or rate) of any such slowing is. All rates of translation inferred in Chapter II are relative to the ribosomal

coverage observed along a specific transcript prior to the encoded slowing feature. What the area under the curve analysis (Chapter II) means as regards the speed per unit time at which different codons are translated is not known. Further, the ribosomal coverage just prior to the encoded focal feature will on average be dependent on expression level. Could the maximum possible ratio (r_{pos}/r_{prec30}) of slowing vary with expression level (proteins produced per mRNA)? For example, in some instances the speed of crowded ribosomes along a transcript might be constrained more by the movement of closely-packed neighboring ribosomes than by encoded slowing features. Such a question is a reminder of the difficulties in inferring what static footprints mean for the activity of ribosomes in a living cell.

Other mechanisms of slowing also remain to be investigated. The hypothesis investigated in Chapter II is whether a ribosome, positioned over a codon awaiting a tRNA, will wait for longer the rarer the tRNA is. An alternative slowing hypothesis revolves around the wobble mechanism. It is known that at least in simple cases where one tRNA recognizes two codons, one via a wobble base-pairing, the non-wobble base pair codon tends to be preferred on the average within a genome (Sharp et al. 2005; Higgs and Ran 2008). However the consequences of this finding in terms of elongation rate are unclear. Although it is sometimes presumed that the wobble pairing is the faster of the two (see e.g. Higgs and Ran 2008), the non-wobble pairing has been found to sometimes be faster, sometimes not differ, and sometimes be slower than the wobble pairing, depending on the identities of the codon, tRNA, and whether the tRNA anticodon has been modified to an alternative base such as queuosine or inosine (Curran and Yarus 1989). And although the non-wobble binding tends to be preferred, this is by no means a guarantee that the preferred non-wobble codon is common when considered amongst the group of all possible codons, or conversely that the unpreferred i.e. wobble codon is globally rare. In other words, we should not expect that the results of an investigation into whether wobble mechanisms slow necessarily correlate with the findings presented in Chapter II regarding rare codons. Could a wobble mechanism then account for appreciable slowing? If so, what proportion of the variance in translation speeds along all transcripts a wobble mechanism could explain will need to be examined.

Does the finding that codons do not slow elongation have any practical impact? Codon usage is widely held to be important in gene design, as changes in synonymous coding content can have a great effect on the ultimate titres of functional protein (e.g. Levy et al. 1996). Why should this be so if we find no variation in elongation speed between codons? Several possible explanations exist. Changes to mRNA structure and may alter transcript half-lives and hence the transcript levels of engineered constructs in the cell relative to their wild type counterparts (Stanssens et al. 1986; Hoekema et al. 1987; Petersen 1987). In such a scenario, changes in protein levels which

appear consistent with altered translational efficiency of the transcript may in fact be due to changes in transcript levels or, possibly, translation initiation rates (Kudla et al. 2009). Alternatively, the abnormality of the transgene systems involved may again offer an explanation. Could for example elongation stall so much under altered (high codon demand, low tRNA supply) conditions that elongation could indeed become the rate-limiting step in protein production, at least on certain transcripts? The answer to this is currently unknown. As a start, it has been suggested that under abnormally high levels of protein production, such as often occurs in engineered cells, the translational system may be maxed out to the point that the most efficient codons for translation become those whose tRNAs are predicted to be more rapidly charged under amino acid starvation conditions (Elf et al. 2003; Welch et al. 2009), although this has not been directly shown.

Nor is the picture so simple as presented thus far. Codon usage bias is also postulated to be a function of translational accuracy, with preferred codons selected for to reduce the rates of missense and nonsense errors. Nonsense errors may result in not just a lack of the needed protein, but also production of potentially toxic partial peptides. Missense errors on the other hand can lead to loss of protein function, protein misfolding and aggregation, which in turn may drain cellular sources by sinking energy into the production of useless proteins along with the energy required for the cell to activate a stress response (Drummond and Wilke 2009). The major evidence for the accuracy hypothesis stems from the observation that codon usage bias is often strongest in sites which are evolutionarily conserved and hence likely structurally and functionally important (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008). By what mechanism might codon usage affect accuracy? A ribosome paused awaiting a rare tRNA may be more likely to incorporate the wrong amino acid into the protein as non-cognate tRNAs will have more opportunities to enter the ribosomal A-site, and discrimination and proofreading of correct tRNAs is not perfect. In line with this, experimental alterations of tRNA concentrations can affect accuracy (Kramer and Farabaugh 2007) and in *E. coli*, the frequency of missense errors is diminished by nine-fold if an amino acid is translated by a codon that corresponds to an abundant tRNA rather than a low-abundance one (Precup and Parker 1987).

Thus efficiency and accuracy hypotheses are not necessarily mutually exclusive, and indeed a trade-off is thought to occur between the two (Dong and Kurland 1995; Johansson et al. 2012) such that it is likely impossible to maximize both at the same time. We presume then that a compromise state has evolved between these two parameters to enable 'normal' cell growth. However, the reasons for such co-adaptation are not clear. Understanding the role codons play in translational efficiency versus accuracy is an outstanding issue, and most work, including the analysis presented herein, addresses just one or the other of these hypotheses. Further work

should help delineate to what extent codon usage participates in both efficiency and accuracy. For example, what happens to these two quantities when the normal tRNA:codon proportionality is disturbed, as in transgenic cells? Which one is costlier—or are they both—and why? Does, for example, decreased translational efficiency impede translation of protein products or cause mRNA degradation? Or is initiation always rate-limiting, even under highly stressed experimental conditions where overexpression of a transcript causes an abnormally disproportionate requirement for a rare tRNA, and instead a great majority of fitness defects occur to changes in translational accuracy? Conversely, are cells translationally robust against most mutations if near-cognate tRNAs are likely to have physiochemically similar amino acids?

Thus a number of difficult questions regarding translational selection remain to be answered. However, they are worth asking, as an eventual understanding why tRNA content and codon usage are under co-selection will be relevant both practically and academically. Firstly, most studies on codon usage bias are done under systemically altered in vitro conditions and results then applied to in vivo organisms, but as we have seen this approach may not always be correct. Understanding how stresses to this co-adaption alter experimental results should aid experimental design and interpretation. Secondly, understanding how changes in tRNA concentration effect translation is of practical importance in disease. Both HIV and the host, for example, fight to control changes in tRNA concentrations to skew the tRNA pools toward the codon usage of their respective genes (van Weringh et al. 2011; Li et al. 2012), and tRNA concentrations can be dysregulated in favor of cancer-causing gene expression (Pavon-Eternod et al. 2009). Translational selection also occurs within viruses, many bacteriophages having been shown to evolve towards the codon usage bias and tRNA pools of their host (Sharp et al. 1984; Carbone 2008; Lucks et al. 2008) and some viral genomes have been shown to code for tRNA genes despite the immense selective pressure for them to streamline their genomes (Gingold and Pilpel 2011). Why does such co-adaptation happen? Are disease-causing genes selecting primarily for translational efficiency or accuracy? Additionally, the role tRNAs play in translational efficiency and accuracy will likely be important in understanding the impact of transgenes on cells for either gene therapy or pharmaceutical production, and what tRNA/codon systems to use when engineering novel bacteria.

References

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics 136: 927-935.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. Microbiol Rev 54: 198-210.

Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol 9: 675.

Carbone A. 2008. Codon bias is a major factor explaining phage evolution in translationally biased hosts. J Mol Evol 66: 210-223.

Clarke TFt, Clark PL. 2010. Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. BMC Genomics 11: 118.

Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. J Mol Biol 209: 65-77.

Deana A, Belasco JG. 2005. Lost in translation: the influence of ribosomes on bacterial mRNA decay. Genes Dev 19: 2526-2533.

Dong H, Kurland CG. 1995. Ribosome mutants with altered accuracy translate with reduced processivity. J Mol Biol 248: 551-561.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet 10: 715-724.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341-352.

Elf J, Nilsson D, Tenson T, Ehrenberg M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. Science 300: 1718-1722.

Eyre-Walker A, Bulmer M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Research 21: 4599-4603.

Frenkel-Morgenstern M, Danon T, Christian T, Igarashi T, Cohen L, Hou YM, Jensen LJ. 2012. Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Mol Syst Biol 8: 572.

Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. Mol Syst Biol 7: 481.

Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol 25: 2279-2291.

Hoekema A, Kastelein RA, Vasser M, de Boer HA. 1987. Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: experimental approach to study the role of biased codon usage in gene expression. Mol Cell Biol 7: 2914-2924.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151: 389-409.

Johansson M, Zhang J, Ehrenberg M. 2012. Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. Proc Natl Acad Sci U S A 109: 131-136.

Kolmsee T, Hengge R. 2011. Rare codons play a positive role in the expression of the stationary phase sigma factor RpoS (sigma(S)) in Escherichia coli. RNA Biol 8: 913-921.

Kramer EB, Farabaugh PJ. 2007. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA 13: 87-96.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science 324: 255-258.

Laursen BS, Sorensen HP, Mortensen KK, Sperling-Petersen HU. 2005. Initiation of protein synthesis in bacteria. Microbiol Mol Biol Rev 69: 101-123.

Levy JP, Muldoon RR, Zolotukhin S, Link CJ, Jr. 1996. Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. Nat Biotechnol 14: 610-614.

Li M, Kao E, Gao X, Sandig H, Limmer K, Pavon-Eternod M, Jones TE, Landry S, Pan T, Weitzman MD, David M. 2012. Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. Nature 491: 125-128.

Li X, Hirano R, Tagami H, Aiba H. 2006. Protein tagging at rare codons is caused by tmRNA action at the 3' end of nonstop mRNA generated in response to ribosome stalling. RNA 12: 248-255.

Lucks JB, Nelson DR, Kudla GR, Plotkin JB. 2008. Genome landscapes and bacteriophage codon usage. PLoS Comput Biol 4: e1000001.

Merrikh H, Machon C, Grainger WH, Grossman AD, Soultanas P. 2011. Co-directional replication-transcription conflicts lead to replication restart. Nature 470: 554-557.

Mitarai N, Sneppen K, Pedersen S. 2008. Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. J Mol Biol 382: 236-245.

Parmley JL, Huynen MA. 2009. Clustering of Codons with Rare Cognate tRNAs in Human Genes Suggests an Extra Level of Expression Regulation. Plos Genetics 5.

Pavon-Eternod M, Gomes S, Geslain R, Dai Q, Rosner MR, Pan T. 2009. tRNA over-expression in breast cancer and functional consequences. Nucleic Acids Res 37: 7268-7280.

Petersen C. 1987. The functional stability of the lacZ transcript is sensitive towards sequence alterations immediately downstream of the ribosome binding site. Mol Gen Genet 209: 179-187.

Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. J Biol Chem 262: 11351-11355.

Proshkin S, Rahmouni AR, Mironov A, Nudler E. 2010. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science 328: 504-508.

Rocha E. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 10: 393-395.

Rocha EP, Danchin A. 2003. Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377-378.

Rocha EPC. 2008. The Organization of the Bacterial Genome. Annual Review of Genetics 42: 211-233.

Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation. Cell 153: 1589-1601.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucl. Acids Res. 33: 1141-1153.

Sharp PM, Rogers MS, McConnell DJ. 1984. Selection pressures on codon usage in the complete genome of bacteriophage T7. J Mol Evol 21: 150-160.

Srivatsan A, Tehranchi A, MacAlpine DM, Wang JD. 2010. Co-orientation of replication and transcription preserves genome integrity. PLoS Genet 6: e1000810.

Stanssens P, Remaut E, Fiers W. 1986. Inefficient translation initiation causes premature transcription termination in the lacZ gene. Cell 44: 711-718.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: Selection for translational accuracy. Molecular Biology and Evolution 24: 374-381.

Sunohara T, Jojima K, Tagami H, Inada T, Aiba H. 2004. Ribosome stalling during translation elongation induces cleavage of mRNA being translated in Escherichia coli. J Biol Chem 279: 15368-15375.

Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, Emili A, Xie XS. 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science 329: 533-538.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141: 344-354.

Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12: R110.

van Weringh A, Ragonnet-Cronin M, Pranckeviciene E, Pavon-Eternod M, Kleiman L, Xia X. 2011. HIV-1 modulates the tRNA pool to improve translation efficiency. Mol Biol Evol 28: 1827-1834.

Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C. 2009. Design parameters to control synthetic gene expression in Escherichia coli. PLoS One 4: e7002.